sgRNA constraints and genetic limitations for efficient Cas9 genome editing to generate knock-outs

IRMGARD U. HAUSSMANN[1,2], THOMAS C. DIX[1], DAVID W. J. MCQUARRIE[1], VERONICA DEZI[1], ABDULLAH I. HANS[1], ROLAND ARNOLD[3,4,5] AND MATTHIAS SOLLER[1, 4, 5]

[1]School of Biosciences, College of Life and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

[2]College of Life Science, Birmingham City University, Birmingham, B5 3TN, United Kingdom

[3]Institute of Cancer and Genomics Sciences, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

[4]Birmingham Centre for Genome Biology, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

Running title: Constraints for efficient CRISPR-Cas9 mediated genome editing

Key Words: sgRNA/Cas9 optimization, secondary structure, retro-transposition

[5] Corresponding authors

m.soller@bham.ac.uk R.Arnold.2@bham.ac.uk

Tel: +44 121 414 5905

Abstract

A single guide RNA (sgRNA) directs Cas9 nuclease for gene-specific scission of double-stranded DNA. High Cas9 activity is essential for efficient gene editing to generate gene deletions and gene replacements by homologous recombination. However, cleavage efficiency is below 50% for more than half of randomly selected sgRNA sequences in human cell culture screens or model organisms. Here, we used in vitro assays to determine intrinsic molecular parameters for maximal sgRNA activity including correct folding of sgRNAs and Cas9 structural information. From comparison of over 10 data sets, we find that major constraints in sgRNA design originate from maintaining the secondary structure of the sgRNA, sequence context of the seed region, GC context and detrimental motifs, but we also find considerable variation among different prediction tools when applied to different data sets. To aid selection of efficient sgRNAs, we developed web-based PlatinumCRISPr, a sgRNA design tool to evaluate base-pairing and known sequence composition parameters for optimal design of highly efficient sgRNAs for Cas9 genome editing. We applied this tool to select sgRNAs to efficiently generate gene deletions in *Drosophila Ythdc1* and *Ythdf*, that bind to N^6 methylated adenosines (m^6A) in mRNA. However, we discovered, that generating small deletions with sgRNAs and Cas9 leads to ectopic reinsertion of the deleted DNA fragment elsewhere in the genome. These insertions can be removed by standard genetic recombination and chromosome exchange. These new insights into sgRNA design and the mechanisms of CRISPR/Cas9 genome editing advances use of this technique for safer applications in humans.

Introduction

Bacterially derived Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-associated protein 9 (Cas9) from *Streptococcus pyogenes* provides a powerful tool for precise genome editing (Garcia-Doval and Jinek 2017; Jiang and Doudna 2017; Hille et al. 2018; Doudna 2020). To induce double stand-breaks in DNA at desired locations, the DNA scission enzyme Cas9 uses a guide RNA (gRNA) containing a 20 nucleotide complementary sequence to the genomic target site (protospacer), which also requires the protospacer adjacent motif (PAM) at the 3'end, comprised of NGG sequence (whereby N is any nucleotide and G is guanine). In addition to the target complementary sequence (spacer) the gRNA also contains a constant crispr RNA (crRNA) sequence that base-pairs with *trans*-activating crRNA (tracrRNA). Alternatively, a single gRNA (sgRNA) can be used whereby the crRNA is fused to the tracrRNA through an artificial loop (Jinek et al. 2012; Cong et al. 2013).

High efficiency CRISPR-Cas9 mediated DNA scission is essential to generate mutants at high frequency in genetic screens and to provide the resource for efficient homologous recombination directed gene replacements. Large scale analysis of sgRNA efficiencies revealed the whole spectrum of on-target cleavage activities ranging from 0-100% arguing for a number of parameters that need to be correct for high efficiency cleavage leading to models incorporating weighing of features and/or thermodynamics of secondary structures (Hsu et al. 2013; Doench et al. 2014; Gagnon et al. 2014; Ren et al. 2014; Wang et al. 2014; Chari et al. 2015; Farboud and Meyer 2015; Hart et al. 2015; Housden et al. 2015; Moreno-Mateos et al. 2015; Varshney et al. 2015; Xu et al. 2015; Doench et al. 2016; Liu et al. 2016; Abadi et al. 2017; Gandhi et al. 2017; Chuai et al. 2018; Labuhn et al. 2018; Graf et al. 2019; Zhang et al. 2019; Michlits et al. 2020; Sledzinski et al. 2020; Trivedi et al. 2020; Xiang et al. 2021; Riesenberg et al. 2023). These studies identified that sequences with very low (≤35%) or very high (>80%) guanine-cytosine (GC) overall content were less effective indicating a critical aspect for binding energy in target scission. In addition, purines in the six nucleotides 5` to the PAM substantially increased Cas9 cleavage efficiency, while pyrimidines and in particular uridine resulted in a lower efficiency (Ren et al. 2014; Wang et al. 2014; Housden et al. 2015; Graf et al. 2019). The lower efficiency of two uridine preceding the PAM site was further associated with premature termination of RNA Pol III (Graf et al. 2019), which terminates after a stretch of four to six uridines (Gao et al. 2018). Moreover, changes in internal structure of sgRNA has been found associated with low activity (Moreno-Mateos et al. 2015; Thyme et al. 2016; Jensen et al. 2017). Recently, also bioinformatic approaches employing machine-learning have been used to improve prediction of sgRNA cleavage efficiencies (Xiang et al. 2021). These observations have helped to improve the design of sgRNAs to yield higher efficiencies and incorporated into sgRNA design tools, but correlations between predictions and guide activity vary considerably (Labun et al. 2016) (Haeussler et al. 2016; Sledzinski et al. 2020). Accordingly, available rules to predict sgRNAs are currently not sufficient to guarantee high cleavage efficiency and many sgRNA candidates scoring high fail to cleave efficiently (Labun et al. 2016) (Haeussler et al. 2016; Labuhn et al. 2018; Sledzinski et al. 2020). In particular, the impact of sgRNA folding has not yet been analysed in detail and incorporated in web tools for sgRNA design. The x-ray crystal structure of Cas9 bound to sgRNA has been determined and indicates four regions of base-pairing termed tetraloop (tetraloop forms due to fusion of crRNA and tracrRNA) and stem loops 1-3 that might be important for its function (Anders et al. 2014; Nishimasu et al. 2014). This structure revealed many points of close interactions of the folded sgRNA with Cas9, but whether disruptions in the sgRNA structure would impact on Cas9 cleavage efficiency has not systematically been analyzed (Riesenberg et al. 2023). In particular, highly GC-rich guide RNAs could disrupt the rather weak secondary structure of the sgRNA bound by Cas9.

The CRISPR-Cas9 genome editing tool is widely used to generate knock-out mutants by introducing frameshifts. It has been recognized that introducing premature termination codons (PTCs) can induce use of alternative translation initiation sites (Tuladhar et al. 2019). In addition, in CRISPR-Cas9 engineered "knock-outs" of the m^{6}A mRNA methyltransferase METTL3, it has been found that a functional ORF can be restored by altered splicing leaving considerable levels of m^{6}A in mRNA (Poh et al. 2022). Likewise, compensatory responses have been observed involving upregulating genes and as a consequence causing stronger phenotypes than compared to removing a gene entirely (El-Brolosy et al. 2019; Ma et al. 2019). In addition, some genes have dual functions as protein and RNA (Hachet and Ephrussi 2004). Hence, introducing a frameshift will only remove the protein function. Likewise, many non-coding RNAs are present in introns suggesting that their expression is connected to the expression of the host gene (Boivin et al. 2018), but such relationships have not yet been explored comprehensively as they require more sophisticated genome editing (Deveson et al. 2017). Thus, for generating gene knock-outs, removing the transcription start site or the entire gene should be considered.

Initial concerns about CRISPR/Cas9 were about off target cleavage, but changing one nucleotide in the spacer sequence complementary to the target efficiently abrogates activity (Ren et al. 2014). However, from randomly chosen sgRNA sequences, more than half display cleavage efficiencies below 50%, and we experienced complete inactivity in *Drosophila* mutagenesis or modification of plasmids. From analyzing the causes of such inactivity, we discovered that maintaining the secondary structure of sgRNAs essentially contributes to high-efficiency DNA scission of Cas9, particularly in cold-blooded animals. Likewise, excessive base-pairing in the seed region also impacts on Cas9 cleavage efficiency. Accordingly optimal design of sgRNA for high efficiency DNA cleavage of Cas9 requires analysis of sgRNA secondary structure which is aberrant with about 50% of PAM adjacent sequences in *Drosophila* and humans. To facilitate design of optimal sgRNAs, we developed an online tool incorporating all currently known parameters for sgRNA design including correct sgRNA folding (https://platinum-crispr.bham.ac.uk/predict.pl). However, comparison of different sgRNA cleavage efficiency monitoring screens and various efficiency prediction tools reveals considerable variation among different prediction tools when applied to different data sets. We then applied the PlatinumCRISPr tool to identify high efficiency sgRNAs to generate gene deletions in *Drosophila* using existing transposon marker lines. Using these novel techniques we generated deletions of *Ythdc1* and *Ythdf*, that bind to N^6 methylated adenosines (m^{6}A) in mRNA important in development (Dezi et al. 2016; Haussmann et al. 2016; Roignant and Soller 2017; Balacco and Soller 2019; Anreiter et al. 2021). We further discovered, that generating small deletions with sgRNAs and Cas9 leads to ectopic reinsertion of the deleted DNA fragment elsewhere in the genome, but such inserts can be removed by standard genetic recombination and chromosome exchange. Taken together, large scale analysis of sgRNA cleavage efficiencies in screens together with new sgRNA design tools and insights into the mechanisms of CRISPR/Cas9 genome editing will help to develop this technique for safe application in humans.

Results

RNA secondary structure constraints limit sgRNA/Cas9 activity

Although sgRNA/Cas9 can cleave DNA efficiently, the first sgRNAs (L11GC and R13GC) we designed according to previously published guidelines did not cut the *pUC 3GLA Dscam 3-5* reporter we designed to study *Dscam* alternative splicing from introducing mutations by gap repair recombineering (**Fig 1 A and B, Supplementary Fig S1, Supplementary Table S1**) (Hemani and Soller 2012; Haussmann et al. 2019). Similarly, the first sgRNAs flanking the *Drosophila Ythdf* gene, a reader for m^{6}A mRNA methylation did not result in a deletion of the locus based on screening for loss of an RFP-marked transposon even though we had validated the target sequence in the strain used (n=103, **Supplementary Fig S2, , Supplementary Table S1**)(Balacco and Soller 2019). To determine the intrinsic molecular parameters for maximal sgRNA activity, we devised an in vitro assay to test the DNA scission efficiency of these two sgRNA based on in vitro transcribed sgRNAs and commercially available Cas9 using oligonucleotide and plasmid substrates containing matching protospacer sequences followed by a PAM.

It has previously been shown that extending the tetraloop in the constitutive component of sgRNAs constituted by tracrRNA and crRNA enhances cleavage efficiency in vitro using oligonucleotide substrates (**Fig 1C**) (Jinek et al. 2012). Introducing the extended sequence present in tracrRNA and crRNA into sgRNA (L7GCext) did not increase efficiency of cleaving a plasmid at 37ºC, but increasing the temperature to 42ºC enhanced cleavage by L7GCext (**Fig 1B**). Increasing the salt concentration to 200 mM also did not result in enhanced cleavage by L7GC/R3G (**Fig 1D**). In contrast, both L7GC and L7GCext could cleave an oligonucleotide, while R13GC did not (**Fig 1E**). Using this assay also confirms the previous observation that sgRNAs of shorter length will lead to cleavage. In addition, three guanosines introduced at the 5'end of sgRNAs required for efficient in vitro transcription are tolerated (**Fig 1F and 1G**) (Jinek et al. 2012).

The sgRNA scaffold adopts a typical fold when bound to Cas9 consisting of the bulged tetraloop, followed by small loop 1 and the more extended loops 2 and 3, which form a protective 3'end structure (**Fig 1C**) (Nishimasu et al. 2014). The loop2/3 structure does not involve the uridines incorporated for termination of RNA Pol III driven expression from plasmids (**Fig 1C**). When comparing the secondary structures of the four sgRNAs, we noticed that the two well-cutting sgRNAs L7GC or R3G maintained the secondary structure of the constitutive RNA part, while the non-cutting sgRNA L11GC disrupted the structure of the tetraloop (**Fig 1H-K**). The effect of R13GC seems more subtle as it could cut in the oligonucleotide assay suggesting that the repeated bulge structure is the cause for its inefficiency, which is supported by x-ray crystal structure of the Cas9-sgRNA-DNA complex. Here, the bulge structure is recognized by Cas9 where Tyr_{359} base-stacks with G_{43}, that also forms hydrogen bonds with Asp_{364} and Phe_{351} , and Phe_{351} forms a hydrogen bond with A_{42} (Nishimasu et al. 2014). In addition, the sgRNAs initially used for deleting the *Ythdf* gene have a severely disrupted secondary structure (**Supplementary Fig S2**).

When we systematically analysed genome sequences from *Drosophila* or humans for correct folding and activity of sgRNAs using the above parameters, about 50% of sgRNAs (241 from 481 and 503 from 973, respectively) did not fold properly. Also, only about 10-20 % of randomly selected sgRNAs exert high cleavage efficiency suggesting that correct folding could essentially contribute to high cleavage efficiency (Hsu et al. 2013; Doench et al. 2014; Gagnon et al. 2014; Ren et al. 2014; Wang et al. 2014; Chari et al. 2015; Farboud and Meyer 2015; Moreno-Mateos et al. 2015; Doench et al. 2016; Liu et al. 2016; Abadi et al. 2017; Chuai et al. 2018; Labuhn et al. 2018; Graf et al. 2019; Zhang et al. 2019; Michlits et al. 2020; Sledzinski et al. 2020).

When comparing the sequences of the cutting sgRNAs L7GC or R3G with the non-cutting L11GC and R13GC sgRNAs, we further noticed that L11GC and R13GC sgRNAs contained more guanosines, which in RNA can base-pair with C and U. To test if guanosines in the sgRNA limit Cas9 activity we increased their number in R3G to 13 to make sgRNA 13G (**Fig 2A**). For the design of the R13 sgRNA, care was taken not to disrupt the tetraloop, but we noticed the potential to interfere with loop 2 (**Fig 2B**, see below). Intriguingly, sgRNA 13G is not capable of directing

Cas9 cleavage if the target sequence is present in a 3 kb plasmid, but is active with a short oligonucleotide substrate (**Fig 2A-D**). Likewise, adding a restriction enzyme together with sgRNA/Cas9 inhibited Cas9 in cleaving plasmid DNA suggesting that Cas9's ability to scan DNA can be impaired separately from its ability to cleave DNA.

Since the increased number of Gs in sgRNA R13G lead to enhanced base-pairing, we exchanged the Gs with Cs leading to an open structure in the seed region. This sgRNA R13C cleaved the test-plasmid efficiently (**Fig 2C and 2E**). Introducing Cs in the left or right half of sgRNA R13G lead to short stem loops and inefficient cleavage of the test-plasmid (**Fig 2C, 2F and 2G**).

To further test to what extent base-pairing impacts on Cas9 activity, we generated sgRNAs L10ds6G and R10ds6GC, where the proximal or the distal half leads to complementary base-pairing of the gRNA with the constant part, respectively (**Fig 3A and 3B**). Although both sgRNAs supported Cas9 cleavage of oligonucleotide substrate, the R10ds6GC sgRNA base-pairing with the proximal part was mostly inactive in cleaving the plasmid indicating an impaired ability of Cas9 to scan DNA (**Figs 3C and 3D**).

Taken together, these results demonstrate that the structure of the sgRNA is important for efficient Cas9 mediated DNA scission in vitro. Furthermore, high G content and base-pairing in the distal part of the gRNA also impairs DNA scission, while base-pairing in the proximal part is tolerated.

To further substantiate these findings, we analyzed the structures of sgRNAs from previous studies in mammalian cells and *Drosophila* with regard to their cleavage efficiency of previous attempts to define rules for sgRNA cleavage efficiency in vivo (Ren et al. 2014; Graf et al. 2019). Indeed, in 39 sgRNAs designed for use in *Drosophila* reduced cleavage efficiency in nine sgRNAs is associated with disturbances of the sgRNA secondary structure resulting in a cleavage efficiency below 35% (**Supplementary Figs S3 and S4, Supplementary Table S1**)(Ren et al. 2014). Similarly, from 22 sgRNAs designed for use in mammalian cells, 13 had a cleavage efficiency below 35% associated with disturbances of the sgRNA secondary structure (**Supplementary Fig S5, Supplementary Table S1**) (Graf et al. 2019). Similar result were also observed for the efficiency of sgRNAs in honey bees (Roth et al. 2019).

Given the requirement for correct folding of the sgRNA for efficient Cas9 mediated DNA scission, we further examined the x-ray crystal structure of the Cas9-sgRNA-DNA complex to see whether this would provide additional instructions to design sgRNAs (Nishimasu et al. 2014). Indeed, the first two nucleotides, adenosine 51 and 52 (A_{51} and A_{52}, **Supplementary Fig S6 and S7**) after the tetraloop form an aromatic base-stacking interaction with phenylalanine 1105 (Phe_{1105}) of Cas9. Furthermore, these interactions are stabilized by guanosine 62 (G_{62}) forming non-Watson Crick hydrogen bonds with A_{51} A_{52} and Phe_{1105}, and uracil 63 (U_{63}) forms a base-stacking interaction with A_{52}. These interactions indicate that base-pairing of the gRNA with these nucleotides of the constant part of the sgRNA reduce Cas9 activity in Cas9 cleavage assays in vitro and mutagenesis in vivo.

Large scale evaluation of novel sgRNA design parameters

Next, we incorporated all the features previously published and from this study into a bioinformatic sgRNA design tool, https://platinum-crispr.bham.ac.uk/predict.pl (a stand-alone code is available for non-commercial use upon request from the authors) (Hsu et al. 2013; Doench et al. 2014; Gagnon et al. 2014; Ren et al. 2014; Wang et al. 2014; Chari et al. 2015; Farboud and Meyer 2015; Moreno-Mateos et al. 2015; Doench et al. 2016; Liu et al. 2016; Abadi et al. 2017; Chuai et al. 2018; Labuhn et al. 2018; Graf et al. 2019; Zhang et al. 2019; Michlits et al. 2020; Sledzinski et al. 2020). The included features include validation of intact secondary structures (tetraloop, loop 2 and 3, **Figs 1-3**), presence of a tetraloop bulge mimic (**Fig 1K**), self-complementarity of the gRNA (nts

1-20, **Figs 2F and 2G**) (Moreno-Mateos et al. 2015; Thyme et al. 2016; Jensen et al. 2017), GC content in the six nucleotide seed region of the gRNA (nts 15-20, **Supplementary Data 1**, (Ren et al. 2014; Wang et al. 2014; Graf et al. 2019), GC content of the gRNA (nts 1-20, **Supplementary Data 1**, (Ren et al. 2014; Wang et al. 2014; Graf et al. 2019), the UUYY motif (nts 16-20), which results in complete base-pairing (**Fig 3**) and can act as a Pol III termination signal (Gao et al. 2018), the UCYG and CYGR motifs (nts 16-20) associated with lower cleavage efficiency (Graf et al. 2019) and lack of base-pairing of nucleotides 40, 41, 51 and 52 that are engaged in contacts with Cas9 (**Supplementary Fig S6 and S7**).

To identify additional parameters affecting sgRNA cleavage efficiency, we performed a motif analysis among the 35% low scoring sgRNAs for a number of different data sets (Doench et al. 2014; Gagnon et al. 2014; Ren et al. 2014; Wang et al. 2014; Chari et al. 2015; Farboud and Meyer 2015; Hart et al. 2015; Moreno-Mateos et al. 2015; Varshney et al. 2015; Xu et al. 2015; Doench et al. 2016; Gandhi et al. 2017; Xiang et al. 2021) , but we did not find motifs associated with low performance in individual data sets.

We then analysed the performance of a *Drosophila* sgRNA data set (Ren et al. 2014) according to sgRNA design parameters described above. Our design tool PlatinumCRISPr selected 13 from 39 sgRNAs and those showed a cleavage efficiency of 55 % or more (**Fig 4A**). We then analysed a number of sgRNA prediction tools for this data set including Chariscore (Chari et al. 2015), Crispron (Xiang et al. 2021), DeepSpCas9 (Kim et al. 2019), DoenchScore (Doench et al. 2014), Azimuth (implemented in ChopChop)(Doench et al. 2016), Moreno-Mateos Score (Moreno-Mateos et al. 2015), Wang Score (Wang et al. 2014), Wong Score (Wong et al. 2015) and Xu Score (Xu et al. 2015). PlatinumCRISPr significantly outperformed all of these prediction tools with the *Drosophila* data set (Ren et al. 2014)(**Fig 4B**).

Next, we analysed 14 data sets from various organisms (*Drosophila*, zebrafish, sea squirt, worms and cell culture cells,), which determined sgRNA cleavage efficiency for their performance using the PlatinumCRISPr design tool (Doench et al. 2014; Gagnon et al. 2014; Ren et al. 2014; Wang et al. 2014; Chari et al. 2015; Farboud and Meyer 2015; Hart et al. 2015; Moreno-Mateos et al. 2015; Varshney et al. 2015; Xu et al. 2015; Doench et al. 2016; Gandhi et al. 2017; Xiang et al. 2021). For overall performance (**Fig 5**), six data sets yielded significant ($p \leq 0.05$) enrichment of high efficiency performing sgRNAs (Ren et al. 2014; Wang et al. 2014; Chari et al. 2015; Moreno-Mateos et al. 2015; Xiang et al. 2021), and five showed enrichment ($p \leq 0.25$) (Gagnon et al. 2014; Farboud and Meyer 2015; Hart et al. 2015; Varshney et al. 2015; Xu et al. 2015), while two failed to show enrichment for most of the parameters (Doench et al. 2014; Doench et al. 2016). In this analysis, we noticed that structural constraints were significantly more important in cold-blooded organism, where sgRNAs delivery is by injection in the absence of selection in contrast to cell culture cells, where delivery is by transfection and selection for chronic exposure to sgRNAs for up to 10 days before analysis.

When we analysed the performance of PlatinumCRISPr, Chariscore (Chari et al. 2015), Crispron (Xiang et al. 2021), DeepSpCas9 (Kim et al. 2019), Doench Score (Doench et al. 2014), Azimuth (implemented in ChopChop)(Doench et al. 2016), Moreno-Mateos Score (Moreno-Mateos et al. 2015), Wang Score (Wang et al. 2014), Wong Score (Wong et al. 2015) and Xu Score (Xu et al. 2015) prediction tools (**Supplementary Fig 8A-L**) with the different data sets determining sgRNA cleavage efficiency we found that Wang Score performed best on the Doench data set and that PlatinumCRISPr and Moreno-Mateos performed best with *Drosophila* and zebrafish generated data sets, respectively, but we found no single prediction tool that stood out. In addition, in five out of six cell culture generated data sets none of the prediction tools outperformed the others or substantially increased prediction efficiency (**Supplementary Fig 8A-L**).

As part of this analysis, we noticed that the average cleavage efficiency in the analyzed data sets varied substantially (from 20-75% average cleavage efficiency, Sup Fig 8A-L) pointing towards a bias in outcome when testing various prediction tools with different data sets. To identify common patterns among the different prediction tools applied to different data sets in the following analysis we therefor excluded data sets with cleavage efficiencies below 30% for further analysis (Gagnon et al. 2014; Chari et al. 2015; Farboud and Meyer 2015; Doench et al. 2016).

The default element of ChopChop is Azimuth (Labun et al. 2019), but also implements elements from Doench Score, Chari Score, Xu Score and Moreno-Mateos Score. We therefore reasoned that a combination of two prediction tools could result in more reliable selection of high efficiency cleaving sgRNAs. When we combined two prediction tools the highest scoring combination was PlatinumCRISPr together with Wong score with an average cleavage prediction of 61% (Fig 6A) and this combination also outperformed in all the remaining data sets (Fig 6B), but this increased the stringency and only very few sgRNAs were selected.

A *Drosophila* transformation vector for expression of two sgRNAs

Next, we applied these novel sgRNA design rules to generate *Drosophila* gene deletions. For this purpose we generated a new fly transformation vector with a GFP marker (Solomon et al. 2018), that is easier to select than previously generate *vermillion* marked vectors that require a *vermillion* mutant background for transgene identification (Port et al. 2014; Trivedi et al. 2020). This vector expresses two 20 nt sgRNAs from *U6.1* and *U6.3* promoters (Supplementary **Fig S9A-C**), harboring a G as first nucleotide as a requirement for expression from the *U6* promoter (Paule and

White 2000; Ren et al. 2013). This plasmid can be generated by incorporating the two sgRNA sequences in PCR primers for single step cloning into the plasmid, while previously published vectors require two cloning steps or plasmid recombination (Port et al. 2014; Trivedi et al. 2020). This sgRNA vector can then be injected into *Drosophila* expressing Cas9 in the germline for CRISPR to induce mutations. Alternatively, this vector can be used to generate a transgenic line via the attB site using phiC31 integrase mediated transformation. This fly strain is then crossed to a line expressing Cas9 in the germline for generation of the desired genetic lesion. Transgenically provided sgRNA/Cas9 generally results in a higher efficiency, because the sgRNAs are been provided maternally.

Efficient generation of gene deletions by sgRNA/Cas9 using transposon markers

To generate deletions of the YTH protein genes *Ythdc1* and *Ythdf*, which are located on the third chromosome, transposon inserts *Mi{MIC}YT521-B*^{MI02006} and *PBac{SAstopDsRed}*^{LL04081} marked with *GFP* or *RFP*, respectively, were combined with an X-linked *vasCas9* or for *nosCas9* germline expression of Cas9. To allow for detecting loss of the transposon in the YTH protein genes the GFP and RFP markers of the *vasCas9* insert had been removed. These flies were then crossed to the GFP marked sgRNA construct inserted on the 3rd chromosome (**Fig 7A and F**). The sgRNA insert has a weak GFP marker and can generally be distinguished from Mi{MIC} inserts. Females from this cross were mated with males containing *TM3 Sb/TM6 Tb* double-balancers in single crosses to recover the sgRNA induced individual deletions and avoid analysis of clonal events. The male progeny was then screened for loss of the GFP or RFP marker. Males which had lost the marker were detected in 100 % (n=9) of the crosses for *Ythdc1* and in 88 % (n=9) for *Ythdf*, respectively. This is a substantial increase of efficiency over imprecise P-element excision with a frequency of

0.01 to 1% (Soller et al. 2006; Haussmann et al. 2016; Haussmann et al. 2022). In those crosses with loss of markers, all males for *Ythdc1* had lost the GFP marker, while for *Ythdf* the average frequency of marker loss was 42 % (n=8). In the reverse cross for *Ythdf*, the frequency was 0% (n=5) indicating that the low frequency in males is linked to the absence of recombination in males. The identified single marker-less males were then crossed to *TM3 Sb/TM6 Tb* double-balancers to establish a line and analysed with PCR using primers next to the deletion breakpoints yielding a short PCR product (**Fig 7B and G**). A PCR product had been obtained in all lines were the marker had been lost indicating that the expected deletion had indeed been generated. To generate *Ythdc1* excision lines *nosCas9* was used for germline expression of Cas9 as *vasCas9* together with sgRNAs targeting *Ythdc1* resulted in female sterility.

sgRNA/Cas9 scissioned fragments insert elsewhere in the genome

After establishing the lines, we noticed that all lines (n=10) established for the *Ythdc1* deletion did not show the flightless phenotype previously reported (Haussmann et al. 2016). Therefore, we selected four lines for further analysis by RT-PCR from RNA (*Ythdc1* excision lines, **Fig 7C**) or of genomic DNA (*Ythdf* excision lines, **Fig 7H**) of homozygous flies with primers that were within the deletion and also flanked an intron. Unexpectedly, a copy of the gene was still present in all *Ythdc1* and *Ythdf* excision lines analyzed suggesting that the deleted fragment had been inserted elsewhere in the genome.

To remove this ectopic insert(s), positive lines were crossed to *w+* marked deficiencies and out-crossed for two generations. The X and 2nd chromosomes were then exchanged to establish null-mutant lines from single chromosomes for *Ythdc1* and *Ythdf* that were confirmed by RT-PCR to be free of any ectopic inserts (**Fig 7D, E and Fig 7I, J**). To avoid such complications in the future we generated a PBac w+ containing vector which can be efficiently inserted into a locus by cloning left and right homology arms either to induce a partial deletion upon insertion. Alternatively, mutations in sgRNA cleavage sites can be introduced into the homology arms to insert point mutations followed by scar less removal of the PBac w+ by transposase.

Discussion

DNA scission by the sgRNA/Cas9 complex is highly specific and requires complete base-pairing between the sgRNA and the target DNA generally not tolerating single miss-matches (Ren et al. 2014; Farboud and Meyer 2015). This feature makes the sgRNA/Cas9 complex an ideal tool for genome editing, but its use is currently limited by the low predictability to cut its target in the genome (Haeussler et al. 2016; Labuhn et al. 2018; Sledzinski et al. 2020).

Here, we discovered that the structure of the sgRNA is a key determinant for the scission efficiency of the sgRNA/Cas9 complex in *Drosophila.* However, only about 50% of sequences adjacent to PAM sites constitute sgRNAs that fold properly or are not compromised by unfavourable base-pairing. In support of these two levels of interference, sgRNA R13GC correctly folds the tetraloop and loop2/3, but did not cleave a short oligonucleotide substrate. This indicates that the structure of this sgRNA blocks the catalytic activity of the sgRNA/Cas9 complex, likely by mimicking the bulge structure of the tetraloop. Second, some sgRNAs allowed cleavage of short oligonucleotide substrates (e.g. L7GC and R10ds6GC), but did not support efficient DNA scission of the target sequence in the context of a 3 kb test-plasmid. Likely, these sgRNAs interfere with the ability of the sgRNA/Cas9 complex to scan the DNA for target sites.

When we analyzed cleavage efficiencies of sgRNAs used in Drosophila (Ren et al. 2014), we observed a good overlap with the ability of those sgRNAs to adopt the correct structure and having a high cleavage efficiency. Further refinement to the design of sgRNAs comes from the recognition that the GC content in the seed region is a major determinant to cleavage efficiency in addition to general GC content. Analysis of the x-ray crystal structure of the Cas9-sgRNA-DNA complex also revealed that the two As in the tracrRNA before the tetraloop engage with Cas9 through base stacking and hydrogen bonds (Nishimasu et al. 2014). Base-pairing of these two As with Us at the end of the sgRNA (N_{19} and N_{20}) before the start of the tracrRNA impact sgRNA/CAS9 complex function and reduce cleavage efficiency (Graf et al. 2019). Likewise, if sgRNAs are made by RNA Pol III, terminate occurs at the boundary of the tracrRNA if two UU precede the GUUUU of the start of the tracrRNA (Arimbasseri and Maraia 2015; Graf et al. 2019). Motif searches in various dataset to determine sgRNA cleavage efficiencies did not reveal any further motifs that impact on cleavage efficiency. If such bias exists, this would like have been exploited by parasites of prokaryotic hosts.

The bacterial CRISPR-Cas9 system consists of two RNAs (crRNA and trRNA) that assemble with Cas9 to form the active complex. crRNA and trRNA base-pair through sequence complementarity to form the tetra loop in the active Cas9 complex (Garcia-Doval and Jinek 2017; Jiang and Doudna 2017; Hille et al. 2018). In type II systems the tracrRNA is required for crRNA maturation suggesting that base-pairing takes place while being assembled with Cas9. After the crRNA is trimmed, the entire CRISPR-Cas9 complex can scan genomic DNA for DNA scission sites. Alternatively, the crRNA could hybridize first to genomic DNA and recruit tracrRNA and Cas9 to form a complex for DNA scission on site. In this scenario, the crRNA would not be able to interfere with tracrRNA/Cas9 complex activity by forming an aberrant RNA secondary structure, but whether this second scenario could be applied to more efficient genome editing with reduced off-target cleavage needs to be tested. Of note, the effect of structural constraints are stronger in cold-blooded animals likely reflecting that the optimal temperature for *E. coli* is 37ºC. In this context, it would be worth exploring how much longer the sgRNA can be to still support Cas9 DNA scission as longer RNAs would more stably hybridize to DNA. In either case, however, understanding sgRNA/CAS9 complex assembly will inform how to prevent off-target DNA scission.

In this study, we also compared various sgRNA cleavage efficiency prediction tools with 12 datasets that have determined sgRNA cleavage efficiencies in human cells and various model organisms including Drosophila, zebrafish, C. elegans, honey bees and seasquirt (Doench et al. 2014; Gagnon et al. 2014; Ren et al. 2014; Wang et al. 2014; Chari et al. 2015; Farboud and Meyer 2015; Hart et al. 2015; Moreno-Mateos et al. 2015; Varshney et al. 2015; Wong et al. 2015; Xu et al. 2015; Doench et al. 2016; Gandhi et al. 2017; Kim et al. 2019; Roth et al. 2019; Xiang et al. 2021). Here, the PlatinumCRISPr tool deemed best for *Drosophila*, Moreno-Mateos Score for zebra fish and Wang Score for one human cell culture screen, but surprisingly all prediction tools failed to convince for the other five cell culture screens. Possibly, chronic exposure over several days in cell culture systems could lead to a bias in determining sgRNA cleavage, compared to short exposure when injected into early stage embryos like in insects and zebra fish.

Taken together, optimizing sequence composition and structural constraints in sgRNA design essentially contributes to high DNA cleavage efficiency. Accordingly, optimized sgRNAs show very little base-pairing with sequences adjacent to the loop 1 region, or are not complementary to the tetraloop structure and/or the loop2/3 structure. In addition, avoiding base-pairing in the 10 nt seed region prior to the PAM site predicts high efficiency of sgRNAs for DNA scission. Furthermore, avoiding two Us before the PAM site prevents interference with the sgRNA/Cas9 structure. In any case, however, by introducing the target sequence into a plasmid the efficiency of a particular sgRNA can be reliably determined in an in vitro cleavage assay. Although this assay will determine the cleavage efficiency of an sgRNA, cellular features such as chromatin state can also impact on Cas9 mediated DNA scission (Singh et al. 2015). In addition, whether genes are expressed at the time of sgRNA cleavage likely also plays a role in the observed cleavage efficiency as expression is associated with less compacted DNA.

Generating deletions of entire genes, or essential parts of them is the preferred way to generate a null allele. This approach will avoid complications arising from introducing frameshifts at the beginning of the ORF, as translation could reinitiate from later AUG or CUG start codons (Koushika et al. 1999). In addition, in some genes the RNA has functions on its own, as shown for *oscar* RNA that forms a large RNP particle with Oscar and other RNA binding proteins at the posterior pole of a *Drosophila* oocyte (Hachet and Ephrussi 2004; Haussmann et al. 2011). Thus, deletion of the entire gene region will discover such additional functions harbored in gene transcripts.

When generating deletions of entire genes, we were very surprised to discover that the deleted DNA fragment was inserted into the genome and transcribed, resulting in the expression of protein that rescued the flightless phenotype in *Ythdc1* deletion allele. Although the mechanism for the generation of these new inserts is not known, it seemed not to have led to complex chromosomal aberrations, because the inserts could be removed by standard recombination and/or exchange of chromosomes. Retro-transposition has been observed in the *elav* gene, which led to the loss of all introns in *Drosophila* (Samson 2008). Likewise, holometabolous insects generally have three *elav* genes, but honey bees have only one *elav* gene. The honey bee *elav* gene, however, carries features of the other two genes present in *Drosophila*. Hence, *elav* in honey bees could have collapsed in an ancestor from three to one gene by some form of recombination to include parts specific to the other two *elav* genes (Ustaoglu et al. 2021). Likewise, the *amyloid-β precursor protein* (*APP*) gene displays copy number variation in the human brain, which are increased by retro-transposition in sporadic forms of Alzheimer's disease suggesting evolutionary conserved mechanisms for reinsertion of genomic information (Lee et al. 2018). Re-insertion of fragments cut out by the CRISPR-Cas9 system seems not to be specific to *Drosophila* as it has also been observed in human cells (Geng et al. 2022). Since intron loss and gain occurs during evolution, the mechanism underlying insertion of DNA fragments from sgRNA induced deletions might be responsible for these changes. Thus, when generating sgRNA induced gene deletions, it is essential to test for the absence of any transcripts by RT-PCR, but also to validate a deletion at the DNA level by PCR using flanking primers or Southern blots. Alternatively, a GFP cassette with a polyA site can be inserted to terminate the ORF in the beginning, but it needs to be evaluated whether the polyA site in the beginning of the gene is used or whether the exon containing the GFP cassette is skipped (Soller 2006; Wierson et al. 2020). For a more reliable way to generate gene knock-outs we have now developed a PBac w+ marker that can be inserted when generating a deletion. Instead of deleting the entire gene, however, deletion of a 5'part will render it non-functional. In addition, if this 5'part inserts unwantedly, it will be non-functional. In any case, however, a marked locus will allow for rigorous cleaning of the genetic background.

In essence, we have established the rules for designing highly efficient sgRNAs and established methodology to efficiently generate gene deletions. These findings have implications for other RNA based methodologies including prime editing (Anzalone et al. 2019; Bosch et al. 2021).

Materials & Methods

sgRNA/Cas9 directed DNA cleavage

DNA templates for in vitro transcription were reconstituted from synthetic oligonucleotides. As only the T7 promoter needs to be double-stranded for in vitro transcription, a T7 promoter oligonucleotide (CCTGGCTAATACGACTCACTATAG) was annealed to an anti-sense Ultramer (IDT DNA) encoding the entire sgRNA in addition to the T7 promoter. Alternatively, a 60 nt T7 promoter oligonucleotide with a partial sgRNA was annealed to an anti-sense oligonucleotide encoding the tracrRNA (AAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC) for 15 min at 40º C (2 μM) and made double-stranded by extension with Klenow fragment of DNA polymerase I according to the manufacturer's instructions (NEB). Klenow was then heat-inactivated 10 min at 85º C and oligonucleotides were desalted with a G-50 Autoseq Sephadex spin column (GE) before using for in vitro transcription.

Then, sgRNAs were generated by in vitro transcription with T7 polymerase (T7 MEGAscript, Ambion) from synthetic oligonucleotides (0.2 μM) and trace-labeled with ^{32}P alpha-ATP (800 Ci/mmol, 12.5 μM, Perkin Elmer) in a 20 μl reaction according to the manufacturer's instructions. After DNAse I digestion, free nucleotides were removed with a G-50 Probequant Sephadex spin column (GE). Then, sgRNAs were heated for 2 min to 95º C and left at room temperature to adopt folding. Then, sgRNAs were quantified by scintillation counting and analysed on 8-20 % denaturing polyacrylamide gels as described (Dix et al. 2022).

For synthetic substrate DNAs the sense oligonucleotide (1 μM, sgRNA flanking sequences are: TCGAGCATTATATGAAC-sgRNA-GGGTATTGGGGAATTCATTATGC) was labeled with ^{32}P gamma-ATP (6000 Ci/mmol, 25 μM, Perkin Elmer) with PNK (NEB). After heat-inactivation of PNK for 2 min at 95º C, sense and anti-sense (anti-sense sgRNA flanking sequences are: GGCCGCATAATGAATTCCCCAATACCC-as sgRNA-GTTCATATAATGC) oligonucleotides were annealed by letting cool down to room temperature and used in sgRNA/Cas9 cleavage assays. For plasmid sgRNA/Cas9 cleavage assays, these annealed oligonucleotides were cloned into a modified *pBS SK+* using a Xho I and Not I cut vector to assay sgRNA/Cas9 activity.

For sgRNA/Cas9 cleavage assays, DNA/sgRNA/Cas9 ratios of 1/10/10 were used in a 10 µl reaction using the buffer supplied (NEB) and DEPC-treated water (Haussmann et al. 2019). Typically Cas9 (100 nM final) was incubated with sgRNA (100 nM) for 10 min at 25º C before adding oligonucleotides (10 nM final) or plasmid DNA (10 nM, corresponds to ~25 ng/µl final concentration of a 3 kb plasmid). Plasmids were linearized after Cas9 digestion by first heat inactivating Cas9 for 2 min at 95ºC, and then adding 10 µl of a restriction enzyme (5 U) in NEB buffer 3. Adding a restriction enzyme together with Cas9 inhibited DNA scission by Cas9. Cleavage of oligonucleotides was analysed on 8 % denaturing polyacrylamide gels and plasmid DNA was analysed on ethidium bromide stained agarose gels.

RNA secondary structure was analyzed with RNAfold at http://rna.tbi.univie.ac.at (Gruber et al. 2008) using the following tracrRNA sequence: GUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAA AAGUGGCACCGAGUCGGUGCUUUUUU.

Criteria for optimal sgRNA design and Cas9 structural analysis

Criteria to select sgRNAs that maintain the structure required for efficient Cas9 DNA scission were implemented in the server accessible at https://platinum-crispr.bham.ac.uk/predict.pl and are as follows. The first nucleotide of the sgRNA needs to be a G for efficient transcription initiation by RNA Pol III (Paule and White 2000). For T7 mediated in vitro transcription, three Gs need to be added to sgRNAs of 23 nucleotides. Disruption of the tetraloop or loop 2 and 3 structures, and the sequence U A/G G C/U A/G of nucleotides 16-20, which will result in a tetraloop bulge mimic, were classified as low efficiency sgRNAs as these parts are recognized by Cas9 (Nishimasu et al. 2014). Similarly, a hairpin loop in the gRNA consisting of 4 or more base pairing nucleotides, or base pairing of nucleotides 17-20 of the gRNA (U/C U/C U U), or base-pairing of eight nucleotides within the seed region with looped-out nucleotides spaced by three base-pairing nucleotides (N11-N20), or a GC content below 15% (1 or 2 nts) or above 50% (11 or more nts) were also considered low efficiency. Medium efficiency was assigned for gRNAs with a GC content of 15-25 % (3-5 nts) or 40-50% (8-10 nts), or a low CG content in the seed region (less than 5 nts in nucleotides 11-20). In addition, two U's at position 19/20 of the gRNA reduce efficiency because of premature transcription termination (Gao et al. 2018; Graf et al. 2019). Further, we assigned a medium impact if both of the two nucleotides A_{51} A_{52} and G_{62} U_{63} in the three way junction of loop 1 were base paired or seven nucleotides within the seed region (N11-N20) base-paired with looped-out nucleotides spaced by three base-pairing nucleotides. Thus, we deemed an sgRNA optimal to allow Cas9 to cleave DNA with high efficiency, if the GC content is 30-35 % (6-7 nts) and none of the above criteria applied.

Cas9/sgRNA structural complex analysis was done using Chimera as described (Dix et al. 2022).

RNA extraction, RT-PCR and PCR on genomic DNA

Total RNA was extracted using Tri-reagent (SIGMA) and reverse transcribed with Superscript II (Invitrogen) according to the manufacturer's instructions using an oligo dT primer. PCR was done for 40 cycles with 1 µl of cDNA, with 1 µg of genomic DNA or from a single fly after freezing and drying in 200 µl of isopropanol. Primers to detect the sgRNA/Cas9 induced deletion in *Ythdc1* were YT F1 (GCCGCTGTGACGCAGAATTTGTGTG) and YT R1

(GGCCGTGCATGTTGCGCATGTAGTCC), and in *Ythdf* were 64F1 (GCCGAGAAAGTGCACAAGGATACGGAG) and 64R1 (CAAGGAATGGCTGAAGCAGACTCCTTG). Primers to amplify parts of the body of the RNA also flanking an intron were for *Ythdc1* YT F2 (CCACGCTGCCGCAGAACGACGCCAATC) and YT R2 (GCGGCAGATCCAGTCAAGCTCGATGAC), and in *Ythdf* were 64F2 (GAGCTGCCTGTCGATTCCCAACTCGTG) and 64R2 (CCGCCCTCTTCGTGTCGCTCCTTGAAG). Primers to amplify parts of the *ewg* gene have been described elsewhere (Koushika et al. 1999).

Cloning of sgRNAs into *pUC 3GLA U6.1 BbsI*

To clone two sgRNAs expressed by U6 promoters, the "tracrRNA U6.3 promoter" fragment was amplified with left (AAGATATCCGGGTGAACTTC**G**N_{19}GTTTTAGAGCTAGAAATAGC) and right (GCTATTTCTAGCTCTAAAA**C**N_{19}CGACGTTAAATTGAAAATAGG) sgRNA primers from pUC 3GLA U6.1/3 sgRNA using Pwo polymerase (Roche) with initial 30 sec denaturation at 94ºC followed by two cycles 94ºC/30 sec, 49ºC/40 sec, 72ºC/45 sec, then two cycles 94ºC/30 sec, 51ºC/40 sec, 72ºC/45 sec and 22 cycles two cycles 94ºC/30 sec, 56ºC/40 sec, 72ºC/45 sec. Transcription from the U6 promoter initiates with a G (bold, underlined). Although this G does not need to be present in the targeting sequence, it needs be included for folding of the sgRNA. The sequences for the sgRNAs in *Ythdc1* were gACAGGTATTCCCAAACTCAC and GACATGTAGCGTTCCCATGA, and for *Ythdf* were GTCCTGAAATACGAGCACAA and gATAACGAACATGTGGGATCT. The pUC 3GLA U6.1 BbsI vector was cut by BbsI and the "sgRNA1 tracrRNA U6.3 promoter sgRNA2" fragment was cloned by Gibson assembly according to the manufacturer's instructions (NEB). For sequencing, primer U6.1 Fseq

(GCGCGTACGTCCTTCGCATCCTTATG) was used. The sequences for *pUC 3GLA HAi Dscam 3-5, pUC 3GLA U6.1 BbsI* and the *pUC 3GLA U6.1/3 Ythdf sgRNA* have been deposited (MK908409, MK908408 and MK908407).

Drosophila genetics and phiC31 integrase-mediated transgenesis

All *Drosophila melanogaster* strains were reared at 25ºC and 40%–60% humidity on standard cornmeal-agar food in 12:12 h light:dark cycle as described (Haussmann et al. 2013). *CantonS* was used as a wild type control. For the *Ythdc1*, the GFP marked *Mi{MIC}YT521-B*^{MI02006} transposon insert and the *w+* marked *Df(3L)Exel6094* deficiency were used, and for *Ythdf*, the RFP marked *PBac{SAstopDsRed}*^{LL04081} insert and the *w+* marked deficiency *Df(3R)ED6220* were used. For phiC31 mediated transformation, constructs were injected into *y*^{1} *w* M{vas-int.Dm}ZH-2A; PBac{y+-attP-3B}VK00013* with the landing site inserted at 76A as previously described (Haussmann et al. 2013). Prior to insertion of GFP marked constructs, the GFP and RFP markers had been removed from the *y1 w* M{vas-int.Dm}ZH-2A* landing site by Cre mediated recombination (Bischof et al. 2007; Zaharieva et al. 2015).

Implementation of PlatinumCRISPr

PlatinumCRISPr is implemented as a Perl script based web-server iteratively evaluating the rule set described in the main text. A guide is classified as "compromised" if any of the rules is violated. For analysis of an sgRNA consisting the target complementary sequence (spacer) and the constant crispr RNA (crRNA) fused to the tracrRNA through an artificial loop is used whereby the first nucleotide of the 20 nt spacer sequence is a G, because a G is needed for transcription (Jinek et al. 2012; Cong et al. 2013). Folding of the sgRNA is computed using RNAFold (Version 2.4.17) and further processed using bpRNA for subsequent interpretation of the dot-bracket code describing the secondary structure by a custom-made Perl script. Notably, the sequence position is calculated from the 3'end of the tracerRNA to allow for variable length of the spacer (between 18 and 23 nt) for custom applications using synthesized RNA consisting of the spacer fused to the crRNA and hybridized to the tracrRNA.

PlatinumCRISPr classifies guides by a binary outcome and typically reports around 70% of sgRNAs as "compromised". Accordingly, only the top 30% were analysed for the distribution of their reported cleavage efficiency in a given data set for each scoring application (**Supplementary Fig S8).** Statistical significance was calculated using a one-sided Wilcoxon signed-rank test. We are grateful to M. Haeussler for providing published guide sequencing and cleavage efficiency scores (Haeussler et al. 2016). CrisperON and DeepSpCas9 guide sequencing and cleavage efficiency scores were calculated using published web-interfaces (Wong et al. 2015; Xiang et al. 2021).

Acknowledgments

We thank Bloomington and Kyoto stock centers for fly lines, A. Handler for the PiggyBac clone, BacPac for Bac clones and plasmids, the University of Cambridge Department of Genetics Fly Facility for embryo injections for generating transgenic lines, M. Haeusler for datasets and pre-calculated scores, Sung-Eun Yoon from the Korea Drosophila Resource Centre and Ying Tan from Genetivision Corporation for sgRNA efficiency data, Yuan Tian for help with art work and Pawel Grzechnik for comments on the manuscript. For this work we acknowledge funding from the Biotechnology and Biological Science Research Council, the Leverhulme Trust, the MRC DTP to DWJM and a Genetics Society summer studentship to AIH.

Author contributions

IUH, TCD and MS performed biochemistry and genetic experiments, VD constructed vectors and AH performed genetic experiments. RA performed bioinformatics analysis, programming and server implementation. DWJM analysed data. IUH, RA and MS conceived the project and wrote the manuscript with help from all authors.

Declaration of interests

The authors declare no competing interests.

Accession codes

Genbank: MK908409 (*pUC 3GLA HAi Dscam 3-5*), MK908408 (*pUC 3GLA U6.1 BbsI*) and MK908407 (*pUC 3GLA U6.1/3 YTHDF53 sgRNA*)

Supplementary Data

Supplementary Data accompany this paper.

References

Abadi S, Yan WX, Amar D, Mayrose I. 2017. A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. *PLoS Comput Biol* **13**: e1005807.

Anders C, Niewoehner O, Duerst A, Jinek M. 2014. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513**: 569-573.

Anreiter I, Mir Q, Simpson JT, Janga SC, Soller M. 2021. New Twists in Detecting mRNA Modification Dynamics. *Trends Biotechnol* **39**: 72-89.

Anzalone AV, Randolph PB, Davis JR, Sousa AA, Koblan LW, Levy JM, Chen PJ, Wilson C, Newby GA, Raguram A et al. 2019. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**: 149-157.

Arimbasseri AG, Maraia RJ. 2015. Mechanism of Transcription Termination by RNA Polymerase III Utilizes a Non-template Strand Sequence-Specific Signal Element. *Mol Cell* **58**: 1124-1132.

Balacco DL, Soller M. 2019. The m(6)A Writer: Rise of a Machine for Growing Tasks. *Biochemistry* **58**: 363-378.

Bischof J, Maeda RK, Hediger M, Karch F, Basler K. 2007. An optimized transgenesis system for Drosophila using germ-line-specific phiC31 integrases. *Proc Natl Acad Sci U S A* **104**: 3312-3317.

Boivin V, Deschamps-Francoeur G, Scott MS. 2018. Protein coding genes as hosts for noncoding RNA expression. *Semin Cell Dev Biol* **75**: 3-12.

Bosch JA, Birchak G, Perrimon N. 2021. Precise genome engineering in Drosophila using prime editing. *Proc Natl Acad Sci U S A* **118**.

Chari R, Mali P, Moosburner M, Church GM. 2015. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat Methods* **12**: 823-826.

Chuai G, Ma H, Yan J, Chen M, Hong N, Xue D, Zhou C, Zhu C, Chen K, Duan B et al. 2018. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biol* **19**: 80.

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA et al. 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**: 819-823.

Deveson IW, Hardwick SA, Mercer TR, Mattick JS. 2017. The Dimensions, Dynamics, and Relevance of the Mammalian Noncoding Transcriptome. *Trends Genet* **33**: 464-478.

Dezi V, Ivanov C, Haussmann IU, Soller M. 2016. Nucleotide modifications in messenger RNA and their role in development and disease. *Biochem Soc Trans* **44**: 1385-1393.

Dix TC, Haussmann IU, Brivio S, Nallasivan MP, HadzHiev Y, Müller F, Müller B, Pettitt J, Soller M. 2022. CMTr mediated 2'-O-ribose methylation status of cap-adjacent nucleotides across animals. *RNA (New York, NY)* **28**: 1377-1390.

Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R et al. 2016. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol* **34**: 184-191.

Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, Sullender M, Ebert BL, Xavier RJ, Root DE. 2014. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol* **32**: 1262-1267.

Doudna JA. 2020. The promise and challenge of therapeutic genome editing. *Nature* **578**: 229-236.

El-Brolosy MA, Kontarakis Z, Rossi A, Kuenne C, Gunther S, Fukuda N, Kikhi K, Boezio GLM, Takacs CM, Lai SL et al. 2019. Genetic compensation triggered by mutant mRNA degradation. *Nature* **568**: 193-197.

Farboud B, Meyer BJ. 2015. Dramatic enhancement of genome editing by CRISPR/Cas9 through improved guide RNA design. *Genetics* **199**: 959-971.

Gagnon JA, Valen E, Thyme SB, Huang P, Akhmetova L, Pauli A, Montague TG, Zimmerman S, Richter C, Schier AF. 2014. Efficient mutagenesis by Cas9 protein-mediated oligonucleotide insertion and large-scale assessment of single-guide RNAs. *PLoS One* **9**: e98186.

Gandhi S, Haeussler M, Razy-Krajka F, Christiaen L, Stolfi A. 2017. Evaluation and rational design of guide RNAs for efficient CRISPR/Cas9-mediated mutagenesis in Ciona. *Developmental biology* **425**: 8-20.

Gao Z, Herrera-Carrillo E, Berkhout B. 2018. Delineation of the Exact Transcription Termination Signal for Type 3 Polymerase III. *Molecular therapy Nucleic acids* **10**: 36-44.

Garcia-Doval C, Jinek M. 2017. Molecular architectures and mechanisms of Class 2 CRISPR-associated nucleases. *Curr Opin Struct Biol* **47**: 157-166.

Geng K, Merino LG, Wedemann L, Martens A, Sobota M, Sanchez YP, Søndergaard JN, White RJ, Kutter C. 2022. Target-enriched nanopore sequencing and de novo assembly reveals co-occurrences of complex on-target genomic rearrangements induced by CRISPR-Cas9 in human cells. *Genome research* **32**: 1876-1891.

Graf R, Li X, Chu VT, Rajewsky K. 2019. sgRNA Sequence Motifs Blocking Efficient CRISPR/Cas9-Mediated Gene Editing. *Cell Rep* **26**: 1098-1103 e1093.

Gruber AR, Lorenz R, Bernhart SH, Neubock R, Hofacker IL. 2008. The Vienna RNA websuite. *Nucleic Acids Res* **36**: W70-74.

Hachet O, Ephrussi A. 2004. Splicing of oskar RNA in the nucleus is coupled to its cytoplasmic localization. *Nature* **428**: 959-963.

Haeussler M, Schönig K, Eckert H, Eschstruth A, Mianné J, Renaud JB, Schneider-Maunoury S, Shkumatava A, Teboul L, Kent J et al. 2016. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol* **17**: 148.

Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, Mis M, Zimmermann M, Fradet-Turcotte A, Sun S et al. 2015. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**: 1515-1526.

Haussmann IU, Bodi Z, Sanchez-Moran E, Mongan NP, Archer N, Fray RG, Soller M. 2016. m(6)A potentiates Sxl alternative pre-mRNA splicing for robust Drosophila sex determination. *Nature* **540**: 301-304.

Haussmann IU, Hemani Y, Wijesekera T, Dauwalder B, Soller M. 2013. Multiple pathways mediate the sex-peptide-regulated switch in female Drosophila reproductive behaviours. *Proc Biol Sci* **280**: 20131938.

Haussmann IU, Li M, Soller M. 2011. ELAV-mediated 3'-end processing of ewg transcripts is evolutionarily conserved despite sequence degeneration of the ELAV-binding site. *Genetics* **189**: 97-107.

Haussmann IU, Ustaoglu P, Brauer U, Hemani Y, Dix TC, Soller M. 2019. Plasmid-based gap-repair recombineered transgenes reveal a central role for introns in mutually exclusive alternative splicing in Down Syndrome Cell Adhesion Molecule exon 4. *Nucleic Acids Res* **47**: 1389-1403.

Haussmann IU, Wu Y, Nallasivan MP, Archer N, Bodi Z, Hebenstreit D, Waddell S, Fray R, Soller M. 2022. CMTr cap-adjacent 2'-O-ribose mRNA methyltransferases are required for reward learning and mRNA localization to synapses. *Nat Commun* **13**: 1209.

Hemani Y, Soller M. 2012. Mechanisms of Drosophila Dscam mutually exclusive splicing regulation. *Biochem Soc Trans* **40**: 804-809.

Hille F, Richter H, Wong SP, Bratovic M, Ressel S, Charpentier E. 2018. The Biology of CRISPR-Cas: Backward and Forward. *Cell* **172**: 1239-1259.

Housden BE, Valvezan AJ, Kelley C, Sopko R, Hu Y, Roesel C, Lin S, Buckner M, Tao R, Yilmazel B et al. 2015. Identification of potential drug targets for tuberous sclerosis complex by synthetic screens combining CRISPR-based knockouts with RNAi. *Science signaling* **8**: rs9.

Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O et al. 2013. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* **31**: 827-832.

Jensen KT, Floe L, Petersen TS, Huang J, Xu F, Bolund L, Luo Y, Lin L. 2017. Chromatin accessibility and guide sequence secondary structure affect CRISPR-Cas9 gene editing efficiency. *FEBS Lett* **591**: 1892-1901.

Jiang F, Doudna JA. 2017. CRISPR-Cas9 Structures and Mechanisms. *Annu Rev Biophys* **46**: 505-529.

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**: 816-821.

Kim HK, Kim Y, Lee S, Min S, Bae JY, Choi JW, Park J, Jung D, Yoon S, Kim HH. 2019. SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with high generalization performance. *Science advances* **5**: eaax9249.

Koushika SP, Soller M, DeSimone SM, Daub DM, White K. 1999. Differential and inefficient splicing of a broadly expressed Drosophila erect wing transcript results in tissue-specific enrichment of the vital EWG protein isoform. *Mol Cell Biol* **19**: 3998-4007.

Labuhn M, Adams FF, Ng M, Knoess S, Schambach A, Charpentier EM, Schwarzer A, Mateo JL, Klusmann JH, Heckl D. 2018. Refined sgRNA efficacy prediction improves large- and small-scale CRISPR-Cas9 applications. *Nucleic Acids Res* **46**: 1375-1385.

Labun K, Montague TG, Gagnon JA, Thyme SB, Valen E. 2016. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Res* **44**: W272-276.

Labun K, Montague TG, Krause M, Torres Cleuren YN, Tjeldnes H, Valen E. 2019. CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res* **47**: W171-w174.

Lee MH, Siddoway B, Kaeser GE, Segota I, Rivera R, Romanow WJ, Liu CS, Park C, Kennedy G, Long T et al. 2018. Somatic APP gene recombination in Alzheimer's disease and normal neurons. *Nature* **563**: 639-645.

Liu X, Homma A, Sayadi J, Yang S, Ohashi J, Takumi T. 2016. Sequence features associated with the cleavage efficiency of CRISPR/Cas9 system. *Sci Rep* **6**: 19675.

Ma Z, Zhu P, Shi H, Guo L, Zhang Q, Chen Y, Chen S, Zhang Z, Peng J, Chen J. 2019. PTC-bearing mRNA elicits a genetic compensation response via Upf3a and COMPASS components. *Nature* **568**: 259-263.

Michlits G, Jude J, Hinterndorfer M, de Almeida M, Vainorius G, Hubmann M, Neumann T, Schleiffer A, Burkard TR, Fellner M et al. 2020. Multilayered VBC score predicts sgRNAs that efficiently generate loss-of-function alleles. *Nat Methods* **17**: 708-716.

Moreno-Mateos MA, Vejnar CE, Beaudoin JD, Fernandez JP, Mis EK, Khokha MK, Giraldez AJ. 2015. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat Methods* **12**: 982-988.

Nishimasu H, Ran FA, Hsu PD, Konermann S, Shehata SI, Dohmae N, Ishitani R, Zhang F, Nureki O. 2014. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**: 935-949.

Paule MR, White RJ. 2000. Survey and summary: transcription by RNA polymerases I and III. *Nucleic Acids Res* **28**: 1283-1298.

Poh HX, Mirza AH, Pickering BF, Jaffrey SR. 2022. Alternative splicing of METTL3 explains apparently METTL3-independent m6A modifications in mRNA. *PLoS biology* **20**: e3001683.

Port F, Chen HM, Lee T, Bullock SL. 2014. Optimized CRISPR/Cas tools for efficient germline and somatic genome engineering in Drosophila. *Proc Natl Acad Sci U S A* **111**: E2967-2976.

Ren X, Sun J, Housden BE, Hu Y, Roesel C, Lin S, Liu LP, Yang Z, Mao D, Sun L et al. 2013. Optimized gene editing technology for Drosophila melanogaster using germ line-specific Cas9. *Proc Natl Acad Sci U S A* **110**: 19012-19017.

Ren X, Yang Z, Xu J, Sun J, Mao D, Hu Y, Yang SJ, Qiao HH, Wang X, Hu Q et al. 2014. Enhanced specificity and efficiency of the CRISPR/Cas9 system with optimized sgRNA parameters in Drosophila. *Cell Rep* **9**: 1151-1162.

Riesenberg S, Kanis P, Macak D, Wollny D, Düsterhöft D, Kowalewski J, Helmbrecht N, Maricic T, Pääbo S. 2023. Efficient high-precision homology-directed repair-dependent genome editing by HDRobust. *Nat Methods* **20**: 1388-1399.

Roignant JY, Soller M. 2017. m6A in mRNA: An Ancient Mechanism for Fine-Tuning Gene Expression. *Trends Genet* **33**: 380-390.

Roth A, Vleurinck C, Netschitailo O, Bauer V, Otte M, Kaftanoglu O, Page RE, Beye M. 2019. A genetic switch for worker nutrition-mediated traits in honeybees. *PLoS biology* **17**: e3000171.

Samson ML. 2008. Rapid functional diversification in the structurally conserved ELAV family of neuronal RNA binding proteins. *BMC Genomics* **9**: 392.

Singh R, Kuscu C, Quinlan A, Qi Y, Adli M. 2015. Cas9-chromatin binding information enables more accurate CRISPR off-target prediction. *Nucleic Acids Res* **43**: e118.

Sledzinski P, Nowaczyk M, Olejniczak M. 2020. Computational Tools and Resources Supporting CRISPR-Cas Experiments. *Cells* **9**.

Soller M. 2006. Pre-messenger RNA processing and its regulation: a genomic perspective. *Cell Mol Life Sci* **63**: 796-819.

Soller M, Haussmann IU, Hollmann M, Choffat Y, White K, Kubli E, Schäfer MA. 2006. Sex-peptide-regulated female sexual behavior requires a subset of ascending ventral nerve cord neurons. *Current biology : CB* **16**: 1771-1782.

Solomon DA, Stepto A, Au WH, Adachi Y, Diaper DC, Hall R, Rekhi A, Boudi A, Tziortzouda P, Lee YB et al. 2018. A feedback loop between dipeptide-repeat protein, TDP-43 and karyopherin-α mediates C9orf72-related neurodegeneration. *Brain : a journal of neurology* **141**: 2908-2924.

Thyme SB, Akhmetova L, Montague TG, Valen E, Schier AF. 2016. Internal guide RNA interactions interfere with Cas9-mediated cleavage. *Nat Commun* **7**: 11750.

Trivedi D, Cm V, Bisht K, Janardan V, Pandit A, Basak B, H S, Ramesh N, Raghu P. 2020. A genome engineering resource to uncover principles of cellular organization and tissue architecture by lipid signaling. *Elife* **9**.

Tuladhar R, Yeu Y, Tyler Piazza J, Tan Z, Rene Clemenceau J, Wu X, Barrett Q, Herbert J, Mathews DH, Kim J et al. 2019. CRISPR-Cas9-based mutagenesis frequently provokes on-target mRNA misregulation. *Nat Commun* **10**: 4056.

Ustaoglu P, Gill JK, Doubovetzky N, Haussmann IU, Dix TC, Arnold R, Devaud JM, Soller M. 2021. Dynamically expressed single ELAV/Hu orthologue elavl2 of bees is required for learning and memory. *Communications biology* **4**: 1234.

Varshney GK, Pei W, LaFave MC, Idol J, Xu L, Gallardo V, Carrington B, Bishop K, Jones M, Li M et al. 2015. High-throughput gene targeting and phenotyping in zebrafish using CRISPR/Cas9. *Genome research* **25**: 1030-1042.

Wang T, Wei JJ, Sabatini DM, Lander ES. 2014. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**: 80-84.

Wierson WA, Welker JM, Almeida MP, Mann CM, Webster DA, Torrie ME, Weiss TJ, Kambakam S, Vollbrecht MK, Lan M et al. 2020. Efficient targeted integration directed by short homology in zebrafish and mammalian cells. *Elife* **9**.

Wong N, Liu W, Wang X. 2015. WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. *Genome Biol* **16**: 218.

Xiang X, Corsi GI, Anthon C, Qu K, Pan X, Liang X, Han P, Dong Z, Liu L, Zhong J et al. 2021. Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning. *Nat Commun* **12**: 3238.

Xu H, Xiao T, Chen CH, Li W, Meyer CA, Wu Q, Wu D, Cong L, Zhang F, Liu JS et al. 2015. Sequence determinants of improved CRISPR sgRNA design. *Genome research* **25**: 1147-1157.

Zaharieva E, Haussmann IU, Brauer U, Soller M. 2015. Concentration and localization of co-expressed ELAV/Hu proteins control specificity of mRNA processing. *Mol Cell Biol* **35**.

Zhang D, Hurst T, Duan D, Chen SJ. 2019. Unified energetics analysis unravels SpCas9 cleavage activity for optimal gRNA design. *Proc Natl Acad Sci U S A* **116**: 8693-8698.

FIGURE LEGENDS

Figure 1: Sequence dependent in vitro cleavage of oligonucleotides and plasmid DNA by the sgRNA/Cas9 complex.

(A) Sequences of sgRNAs with observed cleavage sites indicated by arrow heads. Small letter guanosines used for in vitro transcription are not present in the target DNA sequence. The seed sequence is indicated by a line at the bottom.

(B) Agarose gel showing Cas9 mediated cleavage of the 11.3 kb Dscam 3-5 plasmid for 24 h with indicated sgRNAs. Plasmids were cut with either Acc65I (lanes 2, 3, 6 and 7) or BspEI (lanes 4, 5, 8 and 9) after Cas9 cleavage. The line at the bottom shows a map of the plasmid with restriction sites indicated. Size markers are EcoRI/HinDIII digested λ DNA of 20 kb, 3.6 kb, 1.9 kb and 0.8 kb.

(C) Structure of the sgRNA scaffold from co-crystallization with Cas9 (Nishimasu et al., 2014). Vertical or horizontal lines indicate Watson-Crick base-pairing, and dots or dashed lines indicate non-Watson-Crick base-pairing. Nucleotides base-pairing in loop 1 are bold. Additional base-pairing found in the tracrRNA-crRNA heterodimer is indicated in the extended scaffold (Jinek et al., 2012).

(D) Agarose gel showing Cas9 mediated cleavage of the 11.3 kb Dscam 3-5 plasmid for 24 h with indicated sgRNAs L7GC and R3G. Plasmids were cut with PstI and NotI. The star denotes incomplete cleavage by NotI and the line at the bottom shows a map of the plasmid with restriction sites indicated. Size markers are EcoRI/HinDIII digested λ DNA of 20 kb, 3.6 kb, 1.9 kb and 0.8 kb.

(E) Denaturing acrylamide gel showing Cas9 mediated cleavage of synthetic oligonucleotides with indicated sgRNAs.

(F) Sequences of sgRNAs with variable length. Small letter guanosines used for in vitro transcription are not present in the target DNA sequence.

(G) Denaturing acrylamide gel showing Cas9 mediated cleavage for 1 h of synthetic oligonucleotides with indicated sgRNAs of variable length.

(H-K) Structure of sgRNAs. Nucleotides base-pairing in loop 1 are bold. Red lines in J and K indicate potential base-pairing with nucleotides in loop 2. The red arrow in J indicates the sequence complementarity leading to a bulge in the tetraloop. The red arrows in K indicate a duplication of the bulge structure present in the tetraloop.

Figure 2: Secondary structure of sgRNAs affects Cas9 mediated cleavage efficiency.

(A) Sequences of sgRNAs. Small letter guanosines used for in vitro transcription are not present in the target DNA sequence.

(B) Agarose gel showing Cas9 mediated cleavage after 6 h of 3 kb pBS SK+ test-plasmids containing the target sequence with indicated sgRNAs. Plasmids were linearized with ScaI as in the control after Cas9 heat inactivation. Size markers are EcoRI/HinDIII digested λ DNA of 20 kb, 3.6 kb, 1.9 kb and 0.8 kb.

(C) Denaturing acrylamide gel showing synthetic oligonucleotides before and after sgRNA/Cas9 mediated cleavage for 1h.

(D-G) Structure of sgRNAs. Nucleotides base-pairing in loop 1 are bold. Red lines in D indicate potential base-pairing with nucleotides in loop 2. Green nucleotides indicate mutations compared to sgRNA R13G.

Figure 3: Base-pairing of sgRNAs in the seed-region blocks Cas9 mediated cleavage of a test plasmid.

(A, B) Structure of sgRNAs. Horizontal red lines in A and B indicate artificially introduced base-pairing with the sgRNA scaffold, and vertical red lines in a indicate potential base-pairing with nucleotides in loop 2. Nucleotides base-pairing in loop 1 are bold.

(C) Denaturing acrylamide gel showing synthetic oligonucleotides before and after sgRNA/Cas9 mediated cleavage for 1h. Note that Cas9 cleavage is heterogeneous.

(D) Agarose gel showing Cas9 mediated cleavage after 6 h of 3 kb pBS SK+ test-plasmids containing the target sequence with indicated sgRNAs. Plasmids were linearized with ScaI as in the control after Cas9 heat inactivation. Size markers are EcoRI/HinDIII digested λ DNA of 20 kb, 3.6 kb, 1.9 kb and 0.8 kb.

Figure 4: PlatinumCRISPr selects high efficiency sgRNAs and outperforms other sgRNA selection tools for a *Drosophila* data set.

(A) Comparison of the in vivo efficiency of all sgRNAs from the Ren et al. (2014) data set with PlatinumCRISPr selected sgRNAs.

B) Comparison of the sgRNA selection performance of PlatinumCRISPr with other sgRNA selection tools.

Figure 5: Comparison of individual sgRNA selection criteria for performance with different data sets. The different data sets with the number of sgRNAs tested are indicated on the left. The different selection criteria are shown on top. Significant and enriched performances are indicated in red and orange, respectively (p<0.05 and p<0.25). Criteria with numbers below 5% are indicated in beige, and criteria already applied to the data set are shown in grey.

Figure 6: Combinations of two sgRNA selection tools select high efficiency cleaving sgRNAs for several sgRNA efficiency screen data sets. Combinations of sgRNA selection tools in A are listed according to overall cleavage efficiency of selected sgRNA significant (p<0.05) for both methods (red) or one method (dark grey). Black indicates events with less than 5% of sgRNAs selected by at least one method. Comparison of sgRNA selection by different sgRNA selection tools shown as

median of the cleavage efficiency for individual data sets (B). The distribution of cleavage efficiencies for all sgRNAs is shown on the left (white box) and for PlatinumCRISPr-Wong Score in red, Wond Score-Xu Score in purple and PlatinumCRISPr-Moreno-Mateos Score in blue.

Figure 7: Generation of gene deletions in *Drosophila* YTH protein genes using two sgRNAs/Cas9 and transposon markers.

(A) Schematic to the *Ythdc1* locus indicating transcripts (white boxes) and the ORF (black boxes) below the chromosome. Primers used are indicated on top and below the transcripts. The w+ marked transposon used for detecting deletions in the locus is indicated by a triangle and the deletion generated is indicated by a line.

(B) Agarose gels showing PCR products amplified from genomic DNA of control or *Ythdc1* transposon excision lines using primers flanking the deletion. Presence of a PCR product indicates the expected gene deletion. DNA markers are indicated on the left.

(C) Agarose gels showing RT-PCR products amplified from cDNA of control or *Ythdc1* transposon excision lines using internal primers flanking an intron. Ectopic insertion in the opposite orientation is indicate by an asterisk (lane 4) as in this instance the transcript is not spliced. DNA markers are indicated on the left.

(D) Agarose gels showing PCR products amplified from genomic DNA of control or *Ythdc1*$^{\Delta 7}$ flies using internal primers and primers flanking the deletion. DNA markers are indicated on the left.

(E) Agarose gels showing RT-PCR products amplified from cDNA of control or *Ythdc1*$^{\Delta 7}$ flies using internal primers in *Ythdf* and *ewg* genes. The PCR product of the *ewg* gene was used as loading control. DNA markers are indicated on the left.

(F) Schematic to the *Ythdf* locus indicating transcripts (white boxes) and the ORF (black boxes) below the chromosome. Primers used are indicated on top and below the transcripts. The RFP marked transposon used for detecting deletions in the locus is indicated by a triangle and the deletion generated is indicated by a line.

(G) Agarose gels showing PCR products amplified from genomic DNA of control or *Ythdf* transposon excision lines using primers flanking the deletion. Presence of a PCR product indicates the expected gene deletion. DNA markers are indicated on the left.

(H) Agarose gels showing PCR products amplified from genomic DNA of control or *Ythdf* transposon excision lines using primers flanking an intron. DNA markers are indicated on the left.

(I) Agarose gels showing PCR products amplified from genomic DNA of control or *Ythdf*$^{\Delta B1}$ flies using internal primers and primers flanking the deletion. DNA markers are indicated on the left.

(J) Agarose gels showing RT-PCR products amplified from cDNA of control or *Ythdf*$^{\Delta B1}$ flies using internal primers in *Ythdf* and *ewg* genes. The PCR product of the *ewg* gene was used as loading control. DNA markers are indicated on the left.

Supplementary Figures

Supplementary Figure S1. Secondary structures of sgRNAs used in Figs 1-3 predicted by RNAfold.

Supplementary Figure S2. Inactive sgRNAs targeted to the *Drosophila ythdf* (*CG6422*) gene.

(A) Sequences of sgRNAs targeting the CG6422 locus for generating a deletion.

(B, C) Secondary structures of sgRNAs. The scaffold is shown on the left indicating Watson-Crick base-pairing by lines and the predicted secondary structure by RNAfold is shown on the right.

Supplementary Figure S3. Secondary structure of sgRNAs targeted to the *Drosophila white* gene from Ren et al. (2014).

(A-AA) Secondary structures of sgRNAs predicted by RNAfold. Minimal energy and proximity base-pair structures are shown on the left and right, respectively. Red and blue indicated high and low probabilities for the adopted structural base-pairing assignment, respectively. The number on the right indicates the effectiveness of inducing heritable mutations after injection into fly embryos as determined by Ren et al. (2014). Red and green numbers indicate an effectiveness which is too high or too low compared to predicted DNA cleavage efficiency. Red arrows point towards structural features likely limiting DNA cleavage and green arrows point towards structural features supporting DNA cleavage.

Supplementary Figure S4. Secondary structure of sgRNAs targeted to the *Drosophila vermillion, ebony* and *yellow* gene from Ren et al. (2014).

(A-L) Secondary structures of sgRNAs predicted by RNAfold targeting the vermillion (A-D), the ebony (E-H) and the yellow gene (I-L) . Minimal energy and proximity structures are shown on the left and right, respectively. Red and blue indicated high and low probabilities for the adopted structural base-pairing assignment, respectively. The number on the right indicates the effectiveness of inducing heritable mutations after injection into fly embryos as determined by Ren et al. (2014). Red and green numbers indicate an effectiveness which is too high or too low compared to predicted DNA cleavage efficiency. Red arrows point towards structural features likely limiting DNA cleavage.

Supplementary Figure S5. Secondary structure of sgRNAs targeted to human CD22 from Graf et al. (2019).

(A-V) Secondary structures of sgRNAs predicted by RNAfold. Minimal energy and proximity base-pair structures are shown on the left and right, respectively. Red and blue indicated high and low probabilities for the adopted structural base-pairing assignment, respectively. The number on the right indicates the effectiveness of inducing heritable mutations after injection into fly embryos as determined by Ren et al. (2014). Red and green numbers indicate an effectiveness which is too high or too low compared to predicted DNA cleavage efficiency. Red arrows point towards structural features likely limiting DNA cleavage.

Supplementary Figure S6. Key sgRNA secondary structural features directly interact with the Spy Cas9 endonuclease in the three way junction .

(A) Minimum Free Energy (MFE) predicted secondary structure of the sgRNA used for co-crystallization of Spy Cas9 (Nishimasu et al., 2014). Nucleotides involved in the three way junction interactions are indicated (A_{51}, A_{52}, G_{62} and U_{63})

(B) X-ray crystal structure of spyCas9 endonuclease in complex with chimeric sgRNA bound to genomic DNA target (PDB: 4OO8) (Nishimasu *et al* 2014). The gRNA portion is coloured in pink and the remainder of the sgRNA is coloured in orange. Genomic DNA is coloured yellow and the protein chain is coloured in green. Nucleotide residues situated at key structural features in the sgRNA are coloured cyan. Note that a small portion of the Cas9 protein chain depicting amino acid side chains to view the sgRNA three-way junction is transparent.

(C) Magnified view of the sgRNA three-way junction. Hydrogen bonds are indicated by black dashed lines and aromatic stacking interactions are shown by pink dashed lines. Here, the interaction of Phe1105 with A_{51}, A_{52} and U_{63} and the interaction of G_{62} with A_{51}, A_{52} and the phosphate backbone can be seen.

Supplementary Figure S7. Key sgRNA secondary structural features directly interact with the Spy Cas9 endonuclease in the tetraloop bulge.

(A) Minimum Free Energy (MFE) predicted secondary structure of the sgRNA used for co-crystallization of spyCas9 (Nishimasu et al., 2014). Nucleotides forming the bulge are indicated (A_{28}, A_{41}, A_{42} and G_{43}).

(B) X-ray crystal structure of Spy Cas9 endonuclease in complex with chimeric sgRNA bound to genomic DNA target (PDB: 4OO8) (Nishimasu *et al* 2014). The gRNA portion is coloured in pink and the remainder of the sgRNA is coloured in orange. Genomic DNA is coloured yellow and the protein chain is coloured in green. Nucleotide residues situated at key structural features in the sgRNA are coloured cyan. Note that the tetra loop after the bulge does not interact with Cas9 and sticks out of the structure.

(C) Magnified view of the sgRNA bulge present in the tetra loop. Hydrogen bonds are indicated by black dashed lines and aromatic stacking interactions are shown by pink dashed lines. Here, the interactions of Phe351, Tyr359 and Asp364 with A_{42} and G_{43} can be seen. Inlet on the right: 180º turn to visualize base-stacking of U_{44} with Tyr325 and His328. Through these interactions base-pairing with G_{27} is prevented.

Supplementary Figure S8. Comparison of sgRNA selection by different sgRNA selection tools shown as median of the cleavage efficiency for individual data sets (A-L). The distribution of cleavage efficiencies for all sgRNAs is shown on the left (white box).

Supplementary Figure S9: *Drosophila* GFP-marked transformation vector for U6 promoter mediated expression of two sgRNAs.

(A) Forward and return primer sequences to incorporate sgRNA sequences. The first nucleotide of the gRNA is indicated by an asterisk.

(B) Cloning scheme for incorporating two sgRNAs into the destination vector.

(C) Plasmid map of the fly transformation vector *pUC 3GLA U6.1/U6.3* sgRNA expressing two sgRNAs under U6.1 and U6.3 promoters, respectively.

(D) Secondary structure of sgRNAs targeting *Ythdc1* and *Ythdf.*

Figure 1

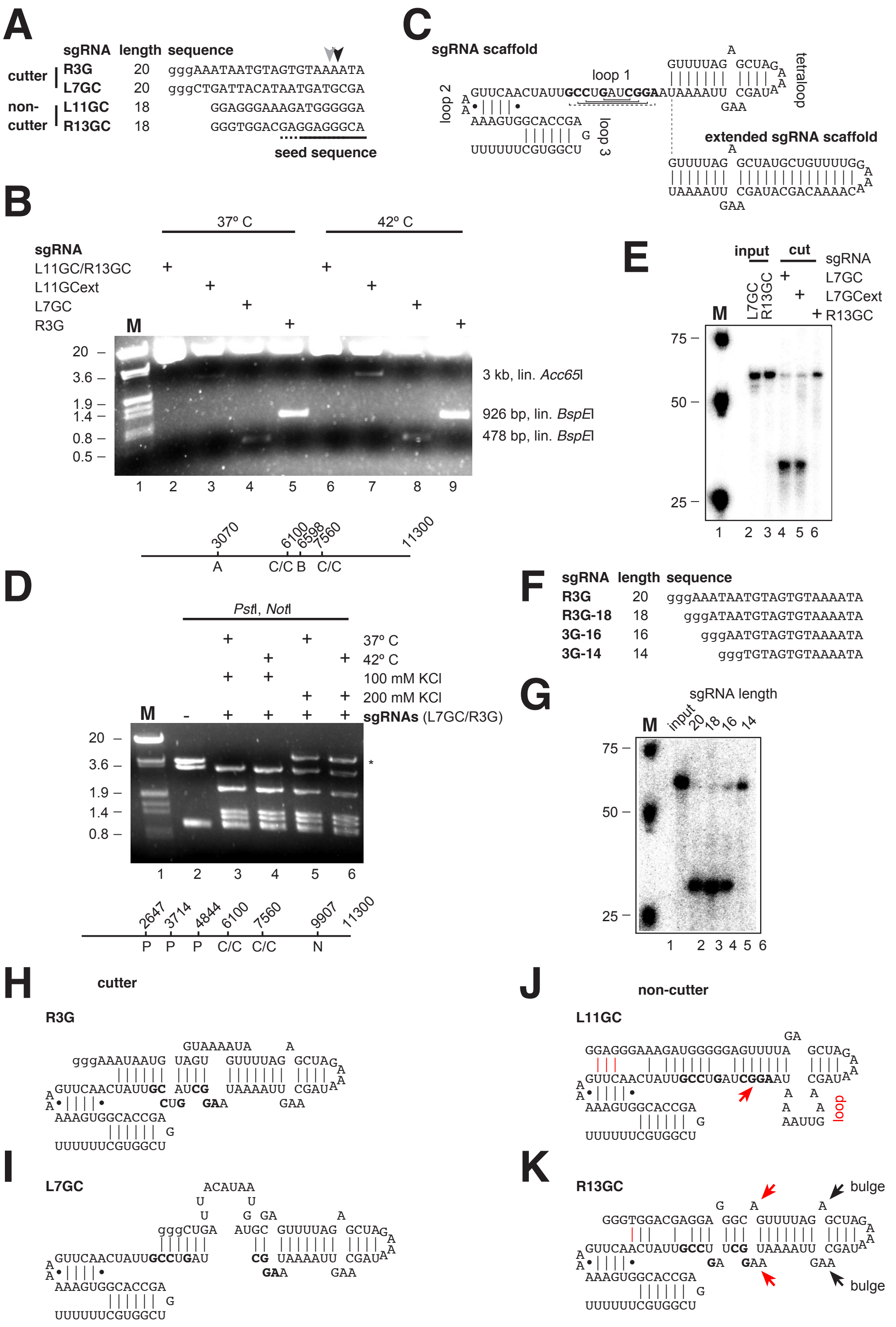
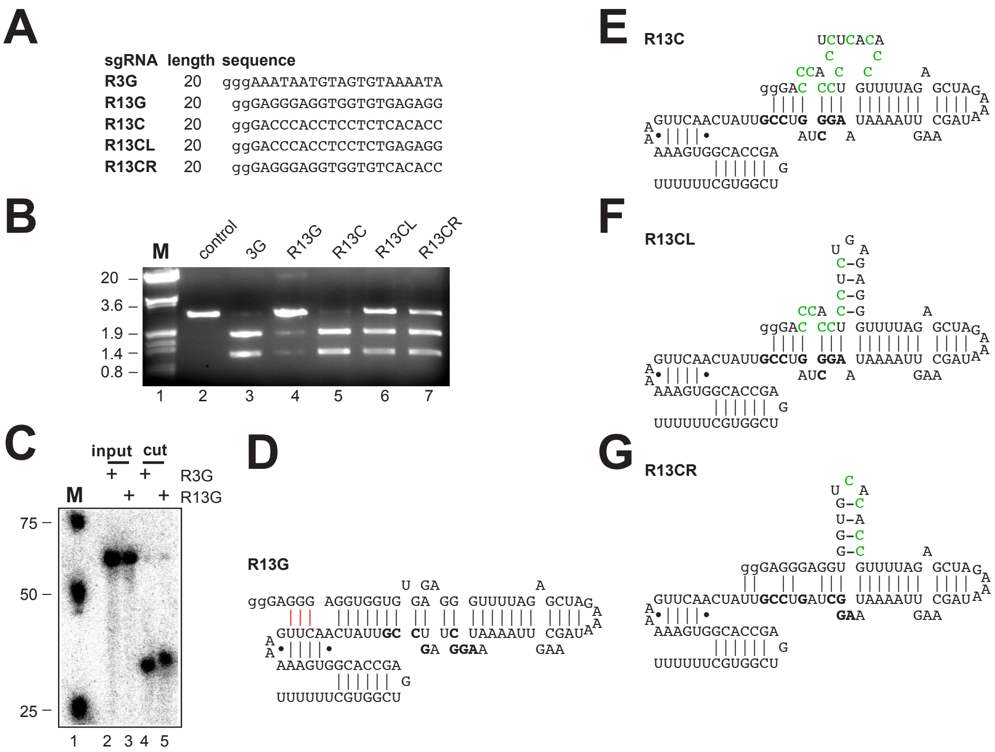

A
sgRNA length sequence
cutter R3G 20 gggAAATAATGTAGTGTAAAATA
L7GC 20 gggCTGATTACATAATGATGCGA
non-cutter L11GC 18 GGAGGGAAAGATGGGGGA
R13GC 18 GGGTGGACGAGGAGGGCA
seed sequence
C
sgRNA scaffold
loop 1
loop 2
loop 3
tetraloop
extended sgRNA scaffold
B
37° C
42° C
sgRNA
L11GC/R13GC
L11GCext
L7GC
R3G
M
20
3.6
1.9
1.4
0.8
0.5
3 kb, lin. Acc65I
926 bp, lin. BspEI
478 bp, lin. BspEI
1 2 3 4 5 6 7 8 9
3070
6100
6598
7560
11300
A
C/C B C/C
E
input
cut
sgRNA
L7GC
R13GC
L7GC
L7GCext
R13GC
M
75
50
25
1 2 3 4 5 6
D
PstI, NotI
37° C
42° C
100 mM KCl
200 mM KCl
sgRNAs (L7GC/R3G)
M
20
3.6
1.9
1.4
0.8
*
1 2 3 4 5 6
2647
3714
4844
6100
7560
9907
11300
P P P C/C C/C N
F
sgRNA length sequence
R3G 20 gggAAATAATGTAGTGTAAAATA
R3G-18 18 gggATAATGTAGTGTAAAATA
3G-16 16 gggAATGTAGTGTAAAATA
3G-14 14 gggTGTAGTGTAAAATA
G
sgRNA length
M
input
20
18
16
14
75
50
25
1 2 3 4 5 6
H
cutter
R3G
I
L7GC
J
non-cutter
L11GC
loop
K
R13GC
bulge
bulge

Figure 2



sgRNA	length	sequence
R3G	20	gggAAATAATGTAGTGTAAAATA
R13G	20	ggGAGGGAGGTGGTGTGAGAGG
R13C	20	ggGACCCACCTCCTCTCACACC
R13CL	20	ggGACCCACCTCCTCTGAGAGG
R13CR	20	ggGAGGGAGGTGGTGTCACACC

Figure 3

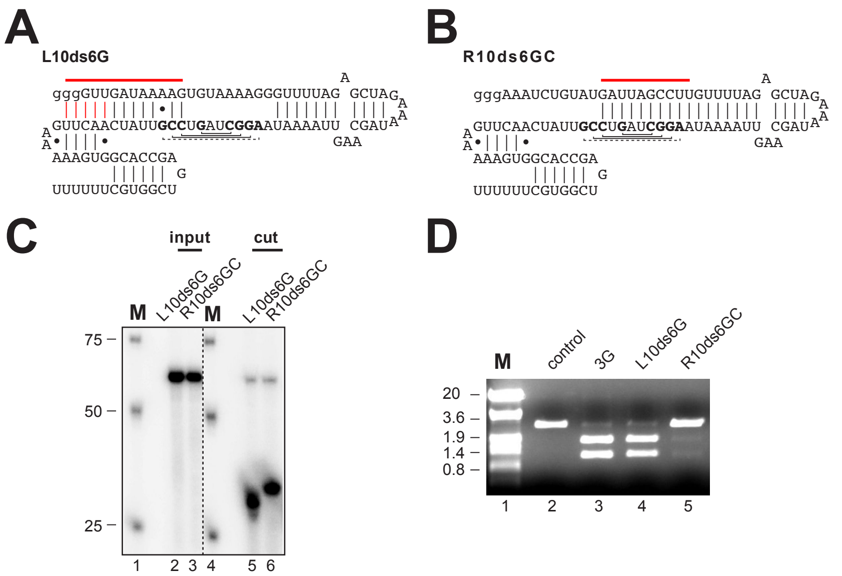

A
L10ds6G
gggGUUGAUAAAAGUGUAAAAGGGUUUUAG GCUAG
GUUCAACUAUUGCCUGAUCGGAAUAAAAAUU CGAU
GAA
AAAGUGGCACCGA
G
UUUUUUCGUGGCU
B
R10ds6GC
gggAAAUCUGUAUGAUUAGCCUUGUUUUAG GCUAG
GUUCAACUAUUGCCUGAUCGGAAUAAAAAUU CGAU
GAA
AAAGUGGCACCGA
G
UUUUUUCGUGGCU
C
input
cut
M
L10ds6G
R10ds6GC
M
L10ds6G
R10ds6GC
75
50
25
1
2
3
4
5
6
D
M
control
3G
L10ds6G
R10ds6GC
20
3.6
1.9
1.4
0.8
1
2
3
4
5

Figure 4

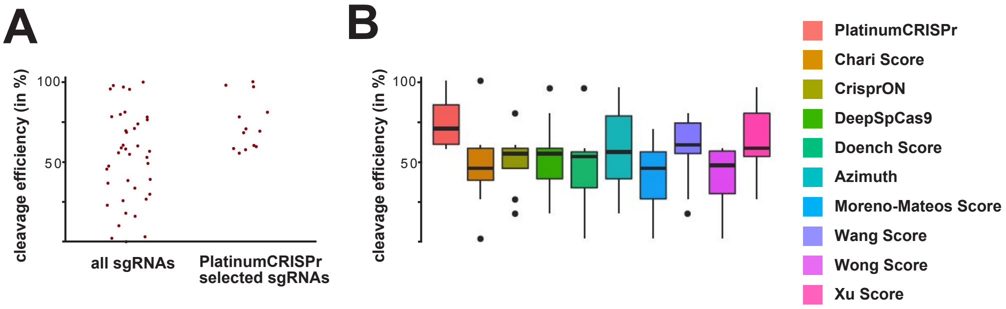

A
B
cleavage efficiency (in %)
100
50
all sgRNAs
PlatinumCRISPr selected sgRNAs
cleavage efficiency (in %)
100
50
PlatinumCRISPr
Chari Score
CrisprON
DeepSpCas9
Doench Score
Azimuth
Moreno-Mateos Score
Wang Score
Wong Score
Xu Score

Figure 5

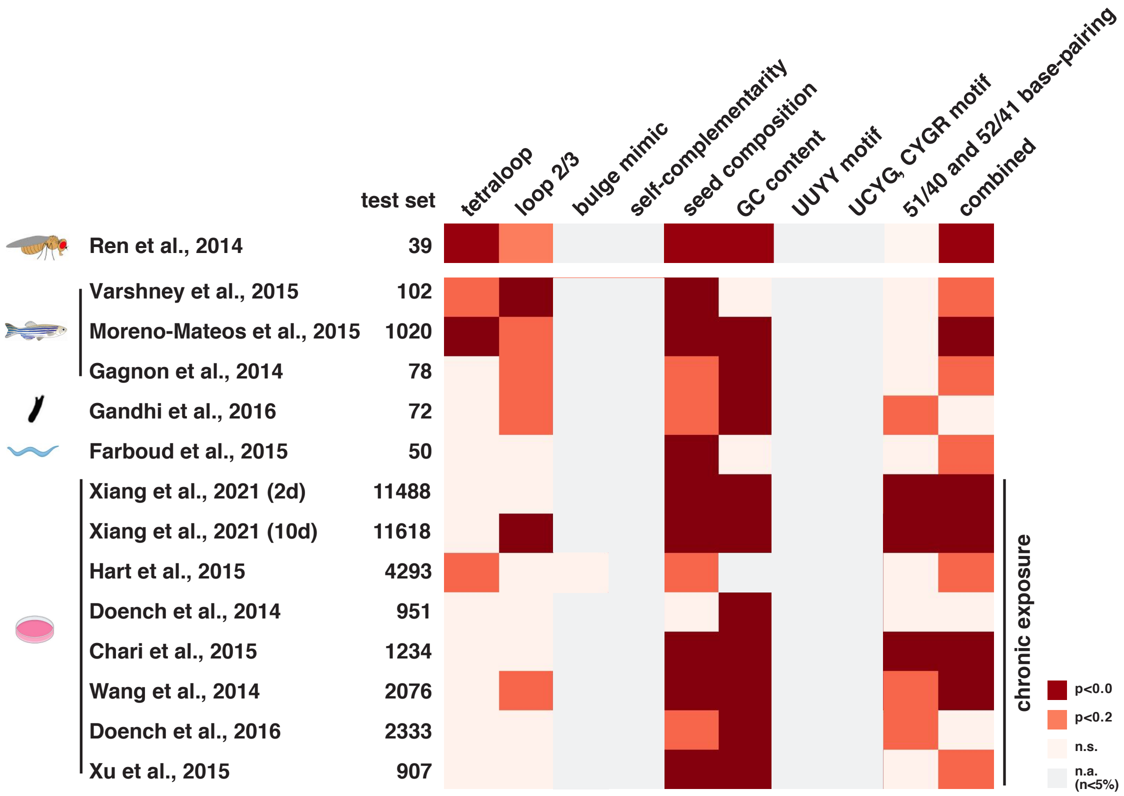
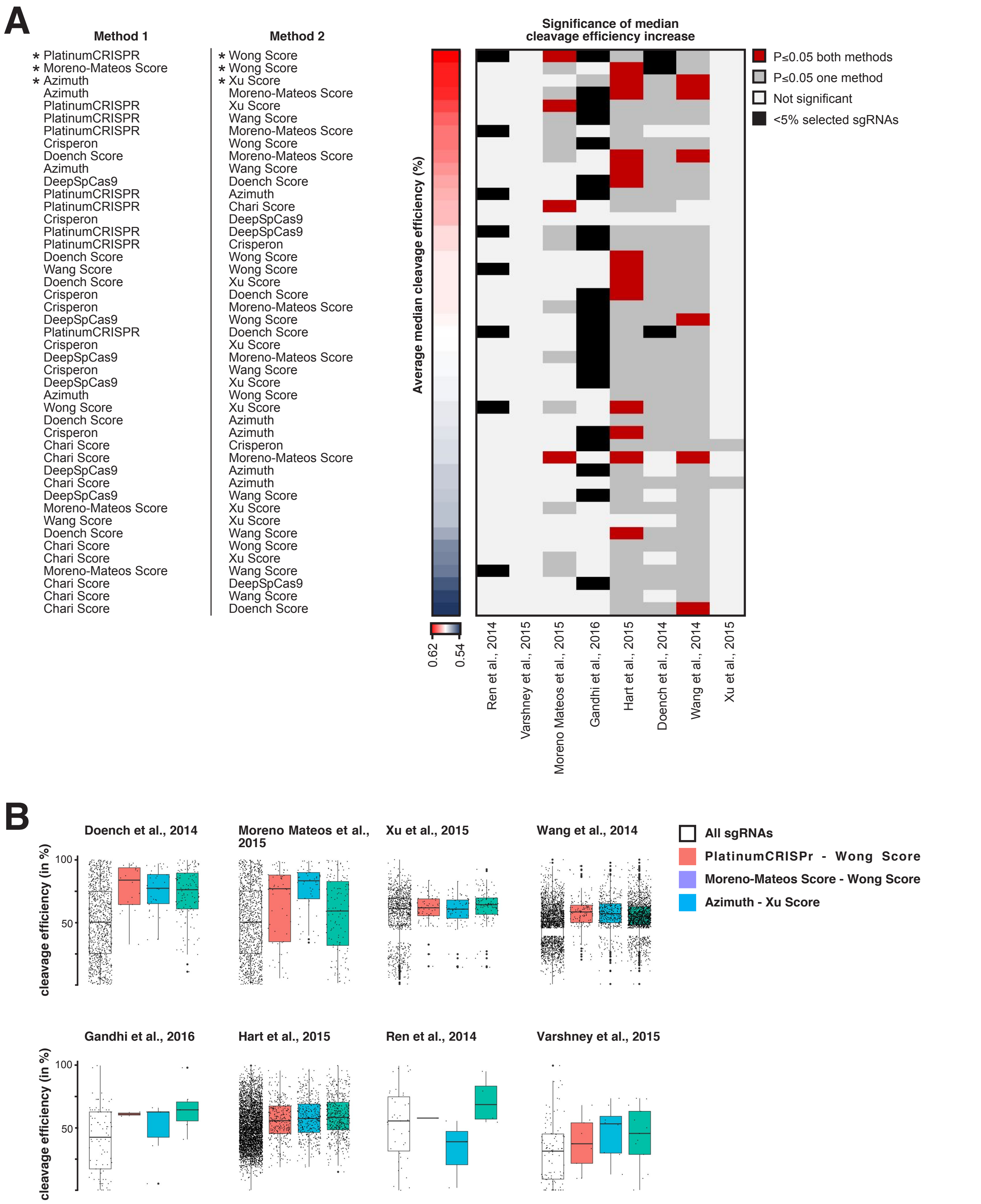
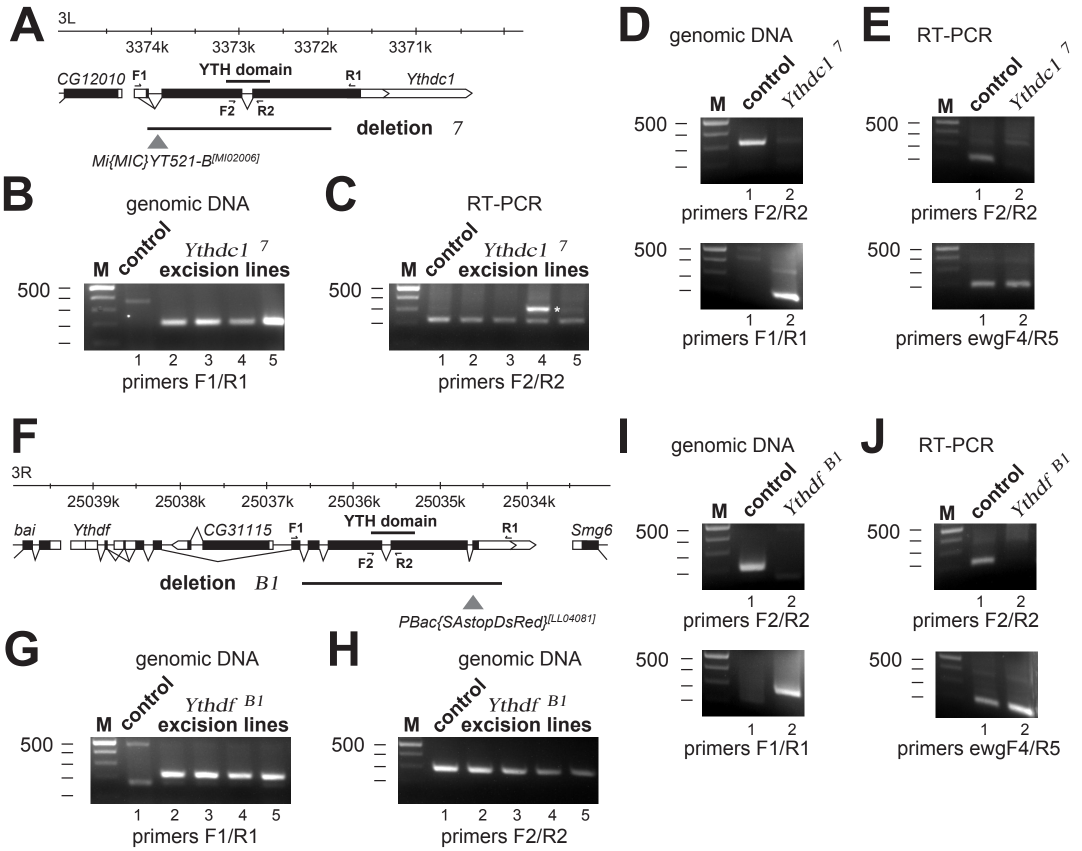

test set
tetraloop
loop 2/3
bulge mimic
self-complementarity
seed composition
GC content
UUYY motif
UCYG, CYGR motif
51/40 and 52/41 base-pairing
combined
Ren et al., 2014
39
Varshney et al., 2015
102
Moreno-Mateos et al., 2015
1020
Gagnon et al., 2014
78
Gandhi et al., 2016
72
Farboud et al., 2015
50
Xiang et al., 2021 (2d)
11488
Xiang et al., 2021 (10d)
11618
Hart et al., 2015
4293
Doench et al., 2014
951
Chari et al., 2015
1234
Wang et al., 2014
2076
Doench et al., 2016
2333
Xu et al., 2015
907
chronic exposure
p<0.0
p<0.2
n.s.
n.a.
(n<5%)

Figure 6

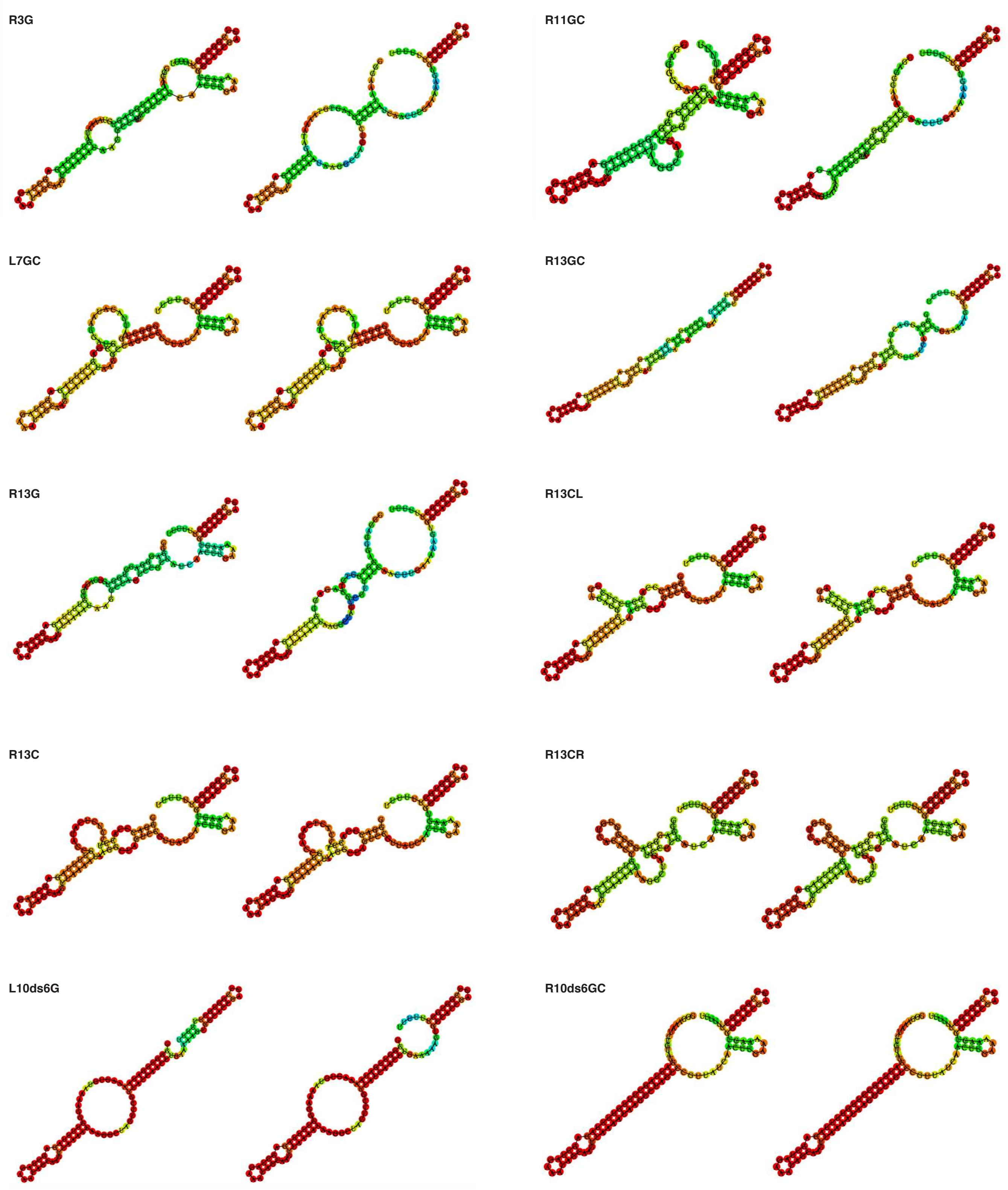


Method 1	Method 2
* PlatinumCRISPR	* Wong Score
* Moreno-Mateos Score	* Wong Score
* Azimuth	* Xu Score
Azimuth	Moreno-Mateos Score
PlatinumCRISPR	Xu Score
PlatinumCRISPR	Wang Score
PlatinumCRISPR	Moreno-Mateos Score
Crisperon	Wong Score
Doench Score	Moreno-Mateos Score
Azimuth	Wang Score
DeepSpCas9	Doench Score
PlatinumCRISPR	Azimuth
PlatinumCRISPR	Chari Score
Crisperon	DeepSpCas9
PlatinumCRISPR	DeepSpCas9
PlatinumCRISPR	Crisperon
Doench Score	Wong Score
Wang Score	Wong Score
Doench Score	Xu Score
Crisperon	Doench Score
Crisperon	Moreno-Mateos Score
DeepSpCas9	Wong Score
PlatinumCRISPR	Doench Score
Crisperon	Xu Score
DeepSpCas9	Moreno-Mateos Score
Crisperon	Wang Score
DeepSpCas9	Xu Score
Azimuth	Wong Score
Wong Score	Xu Score
Doench Score	Azimuth
Crisperon	Azimuth
Chari Score	Crisperon
Chari Score	Moreno-Mateos Score
DeepSpCas9	Azimuth
Chari Score	Azimuth
DeepSpCas9	Wang Score
Moreno-Mateos Score	Xu Score
Wang Score	Xu Score
Doench Score	Wang Score
Chari Score	Wong Score
Chari Score	Xu Score
Moreno-Mateos Score	Wang Score
Chari Score	DeepSpCas9
Chari Score	Wang Score
Chari Score	Doench Score

Figure 7


A
3L
3374k
3373k
3372k
3371k
CG12010
F1
YTH domain
R1
Ythdc1
F2
R2
deletion 7
Mi{MIC}YT521-B[MI02006]
B
genomic DNA
M
control
Ythdc1 7
excision lines
500
1 2 3 4 5
primers F1/R1
C
RT-PCR
M
control
Ythdc1 7
excision lines
500
*
1 2 3 4 5
primers F2/R2
D
genomic DNA
M
control
Ythdc1 7
500
1 2
primers F2/R2
500
1 2
primers F1/R1
E
RT-PCR
M
control
Ythdc1 7
500
1 2
primers F2/R2
500
1 2
primers ewgF4/R5
F
3R
25039k
25038k
25037k
25036k
25035k
25034k
bai
Ythdf
CG31115
F1
YTH domain
R1
Smg6
F2
R2
deletion B1
PBac{SAstopDsRed}[LL04081]
G
genomic DNA
M
control
Ythdf B1
excision lines
500
1 2 3 4 5
primers F1/R1
H
genomic DNA
M
control
Ythdf B1
excision lines
500
1 2 3 4 5
primers F2/R2
I
genomic DNA
M
control
Ythdf B1
500
1 2
primers F2/R2
500
1 2
primers F1/R1
J
RT-PCR
M
control
Ythdf B1
500
1 2
primers F2/R2
500
1 2
primers ewgF4/R5

Supplementary Figure S1



R3G

R11GC

L7GC

R13GC

R13G

R13CL

R13C

R13CR

L10ds6G

R10ds6GC

Supplementary Figure S2

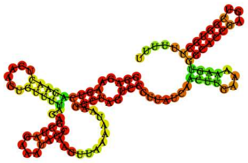
A

sgRNA	length	sequence
CG6422 5'	20	GGAGAAGUCAGAACUCAAGU
CG6422 3'	20	GAGGCAGAAACAAUGGAUCU

B

CG6422 5'

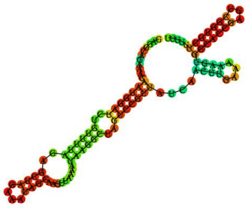
```
                      AA
                     C   G
                     U   U
                      C-G
                      A-U
                      A-U
                      G-U
                 GA   A-U
              GGA  AGUC  AGAGCUAG
              |||  ||||     ||||  A
A GUUCAACUAUUGCCU  UCGG     CGAUA A
   ||||          GA     A   A
A AAAGUGGCACCGA         A   A
       ||||||| G        U   G
  UUUUUUCGUGGCU         A   U
                        A   U
                        A   A
                          A
```



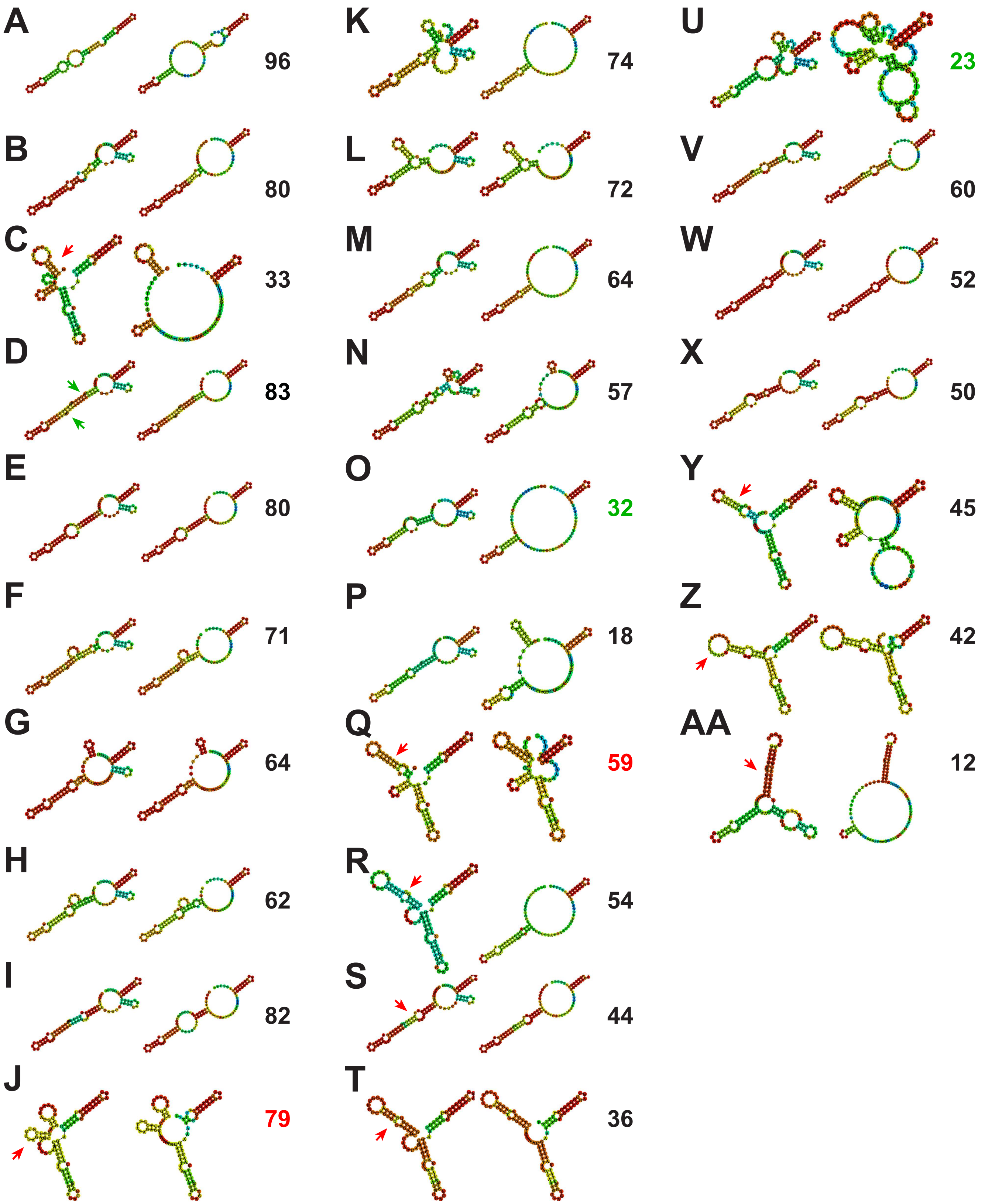
C

CG6422 3'

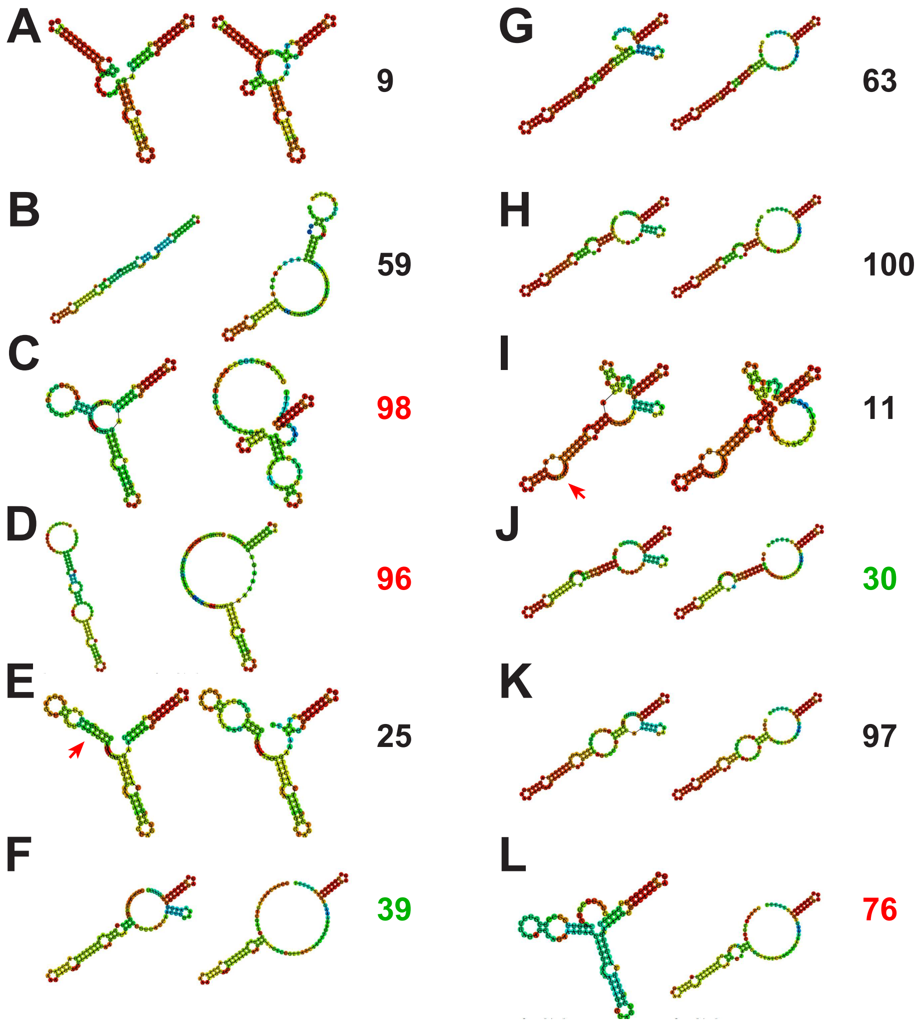
```
                  CU      G A
GAGGCAGAAACAAUGGAU  GUUUUA   GCUAG
           |||||||  ||||||   |||| A
 AGUUCAACUAUUGCCUG  CGGAAU   CGAUA
   ||||           AU      A  A
 AAAAGUGGCACCGA           A  A
       ||||||| G          A  G
  UUUUUUCGUGGCU           A  U
                            U
```



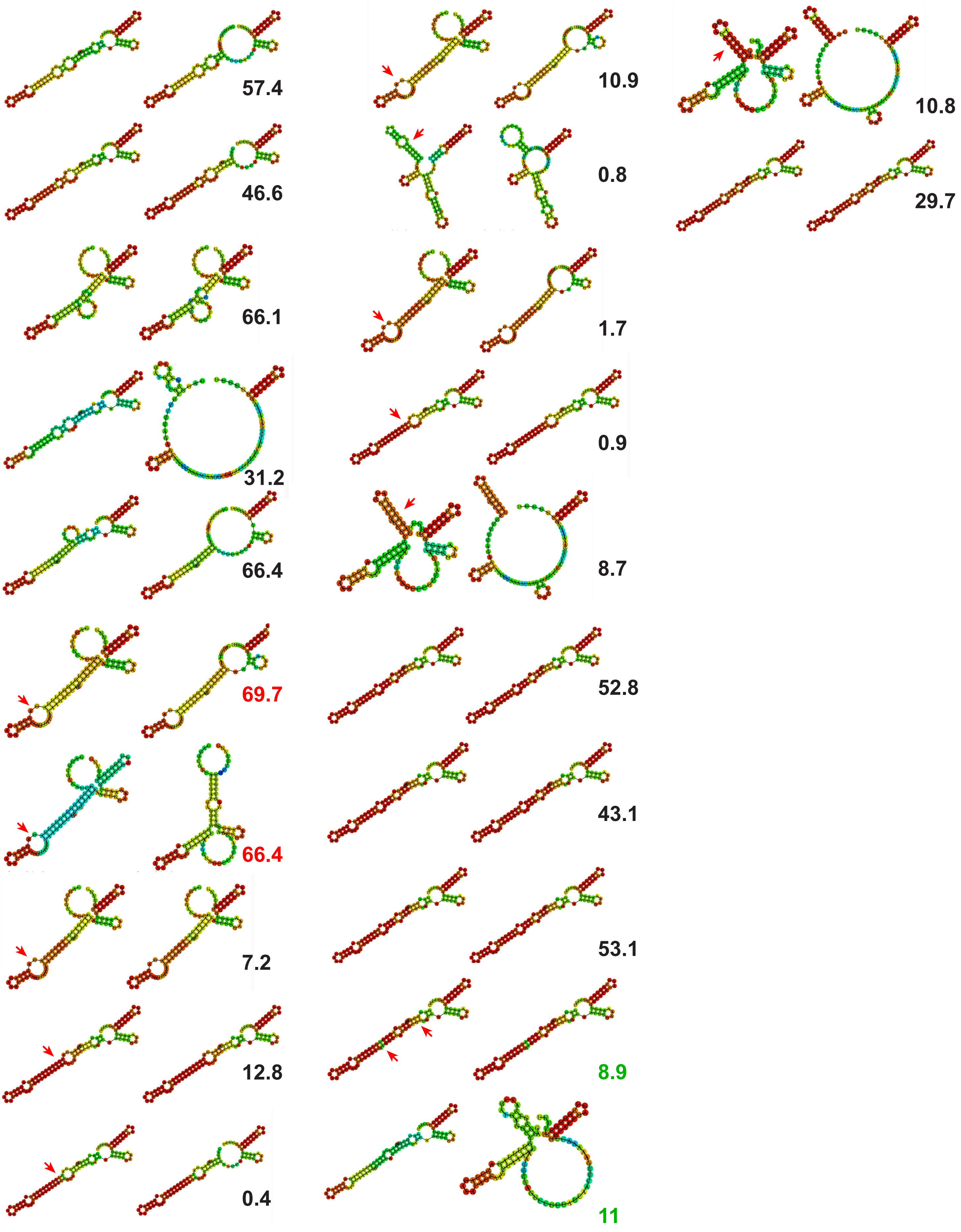
Supplementary Figure S3



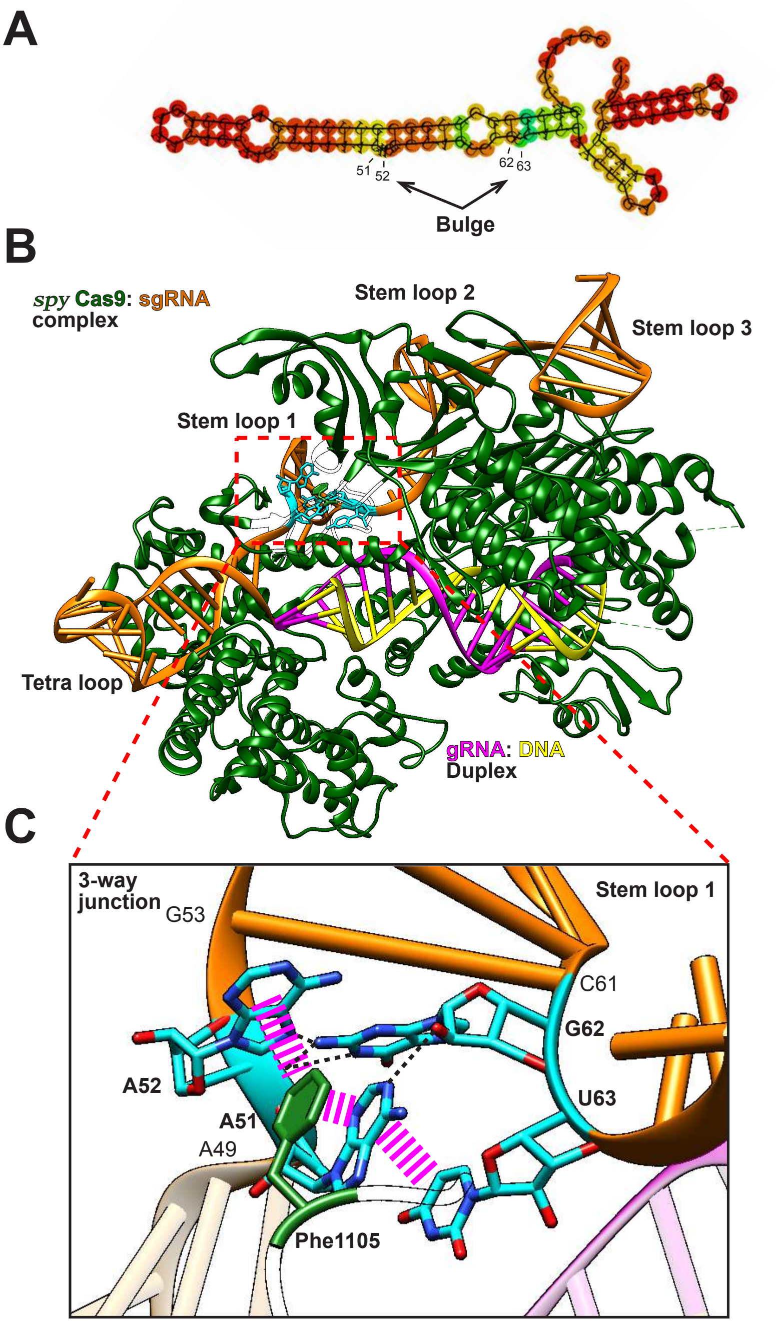
Supplementary Figure S4



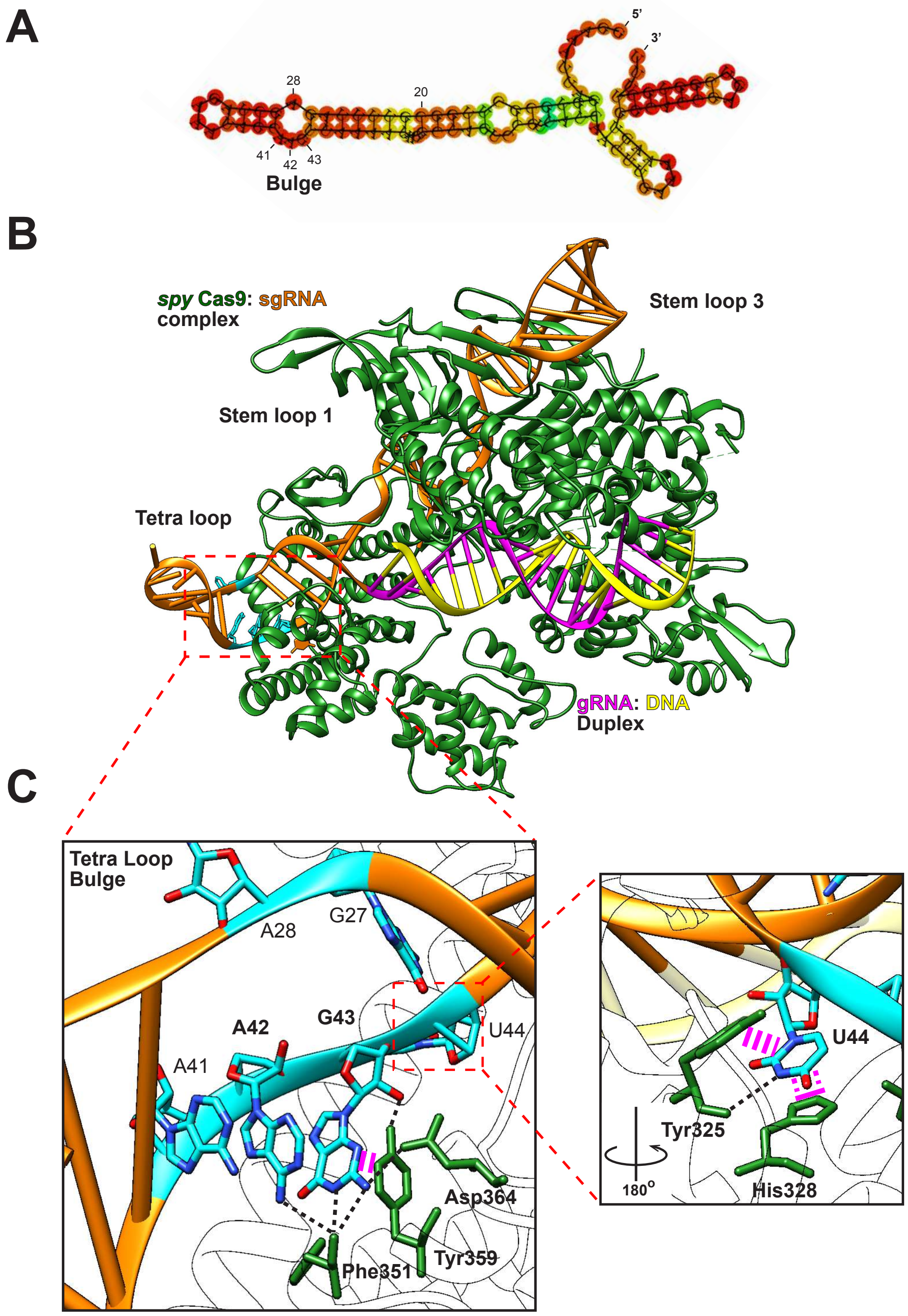
Supplementary Figure S5



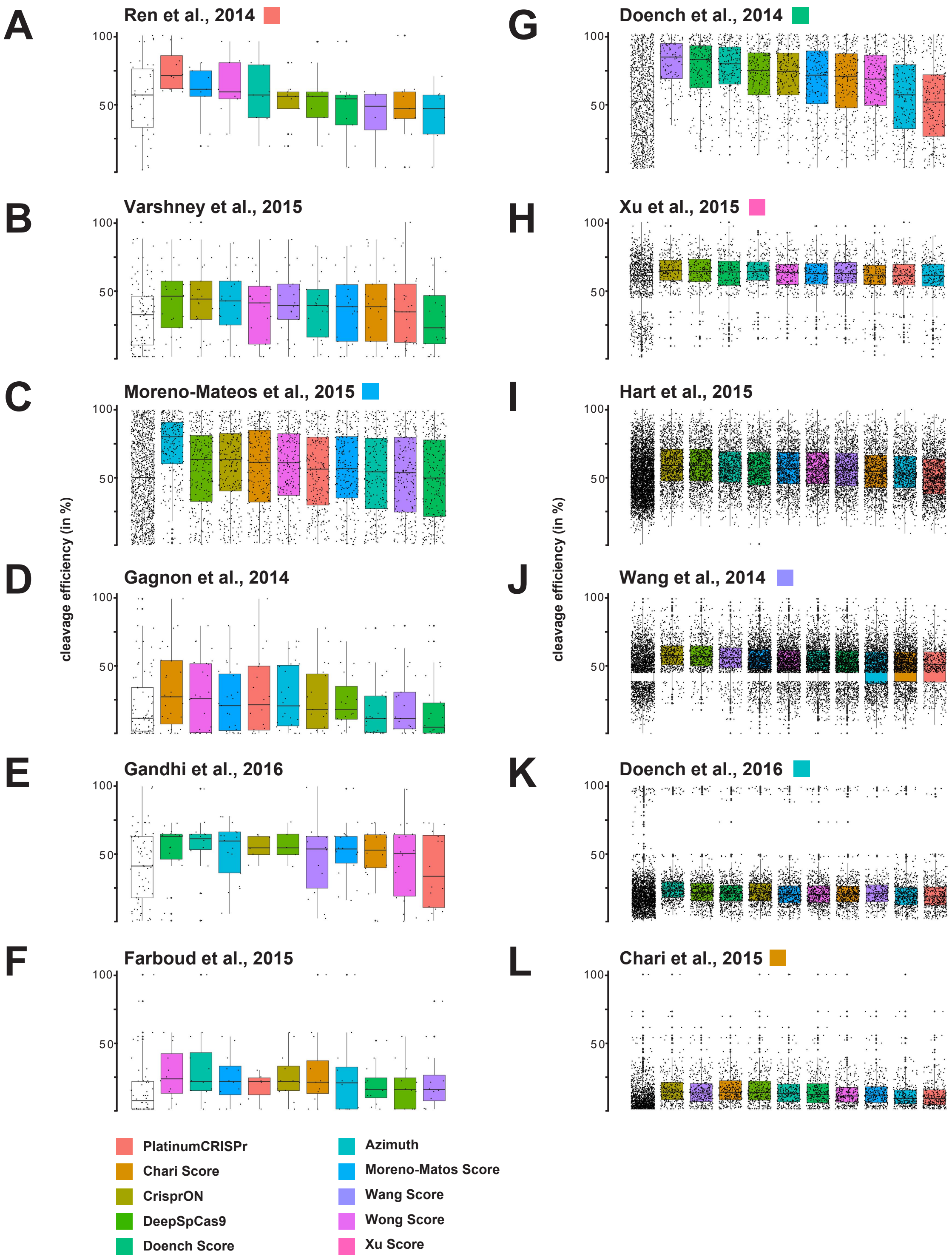
Supplementary Figure S6



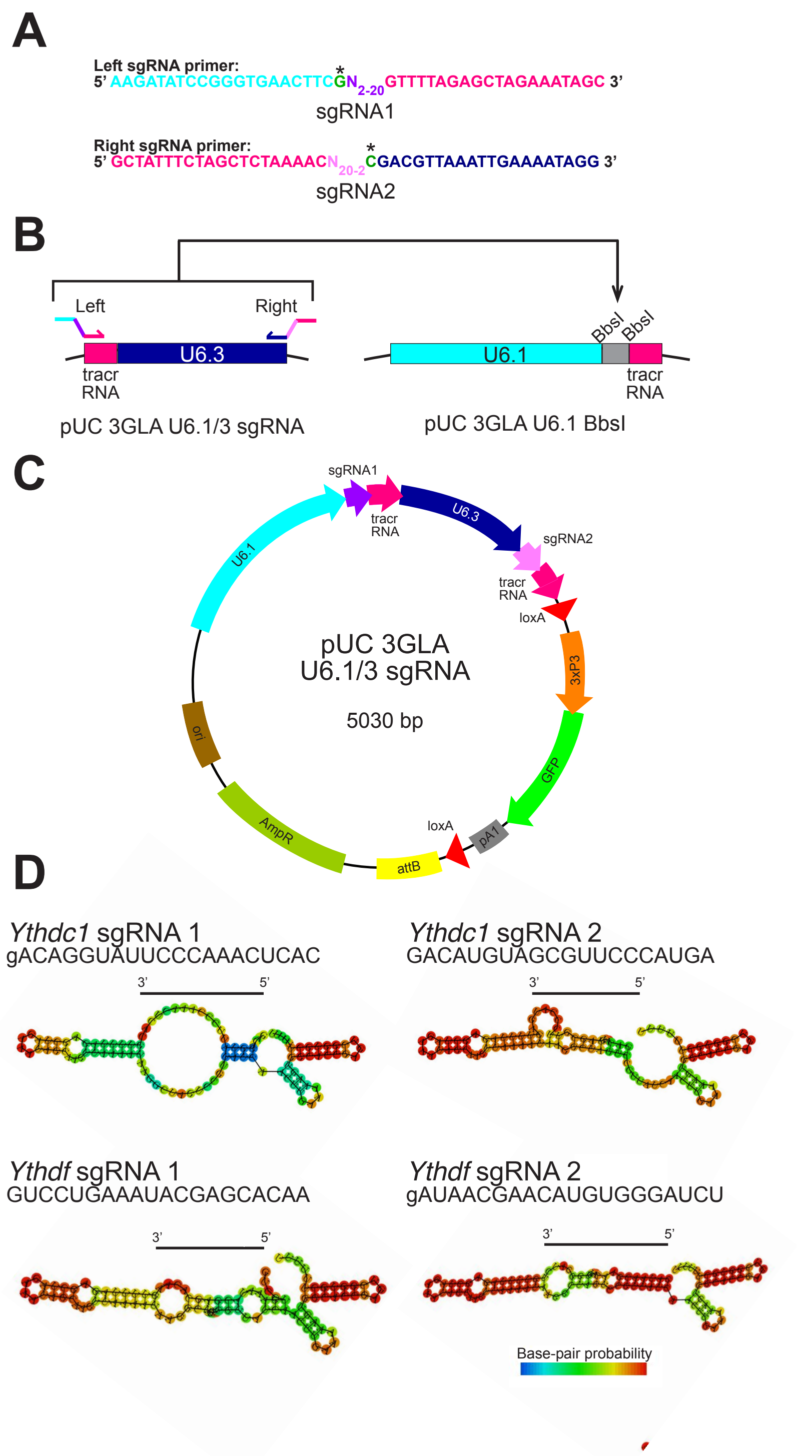
Supplementary Figure S7



Supplementary Figure S8



Supplementary Figure S9


A
Left sgRNA primer:
5’ AAGATATCCGGGTGAACTTCG*N2-20GTTTTAGAGCTAGAAATAGC 3’
sgRNA1
Right sgRNA primer:
5’ GCTATTTCTAGCTCTAAAACN20-2C*GACGTTAAATTGAAAATAGG 3’
sgRNA2
B
Left
Right
U6.3
tracr RNA
pUC 3GLA U6.1/3 sgRNA
U6.1
BbsI
BbsI
tracr RNA
pUC 3GLA U6.1 BbsI
C
sgRNA1
tracr RNA
U6.3
sgRNA2
tracr RNA
loxA
3xP3
GFP
pA1
loxA
attB
AmpR
ori
U6.1
pUC 3GLA U6.1/3 sgRNA
5030 bp
D
Ythdc1 sgRNA 1
gACAGGUAUUCCCAAACUCAC
3’
5’
Ythdc1 sgRNA 2
GACAUGUAGCGUUCCCAUGA
3’
5’
Ythdf sgRNA 1
GUCCUGAAAUACGAGCACAA
3’
5’
Ythdf sgRNA 2
gAUAACGAACAUGUGGGAUCU
3’
5’
Base-pair probability